UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA
Departamento de Filología Románicas, Eslavas y Lingüística General

LA CONSTRUCCIÓN DE TESAUROS ACADÉMICOS: UN MODELO GENERAL Y UN MÉTODO INDUCTIVO CON APLICACIÓN AL "E-LEARNING".

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Ana M. Fernández-Pampillón Cesteros

Bajo la dirección de los doctores

Covadonga López Alonso
Alfredo Fernández-Valmayor Crespo

Madrid, 2010

ISBN: 978-84-693-6551-9

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA

Departamento de Filología Románica, Eslava y Lingüística General

Área de Lingüística General

La construcción de tesauros académicos. Un modelo general y un método inductivo con aplicación al *e-learning*

Memoria para optar al grado de doctor presentada por:

Ana M. Fernández-Pampillón Cesteros

Dirigida por los doctores:

Covadonga López Alonso

Alfredo Fernández-Valmayor Crespo

Noviembre 2009

A mi marido, Miguel
A mis hijos Miguel y Javier

A mi Familia

Agradecimientos

Durante todos estos años de trabajo de tesis he tenido firmes apoyos sin los que difícilmente habría podido empezar y terminarlo. El primero de ellos ha sido el de mis directores, Covadonga López Alonso y Alfredo Fernández-Valmayor, a quienes debo la mayor parte de lo que he aprendido como investigadora y a quienes agradezco no sólo su sabia y experimentada guía sino, también, la confianza permanente y el empuje en los momentos difíciles.

También he contado con el apoyo de mis compañeros y colegas en la Universidad: María Matesanz, compañera de investigación y amiga desde el año 1996, fecha en la que nos incorporamos a la Universidad; le agradezco, sobre todo, sus enseñanzas lexicográficas, lingüísticas y su sereno apoyo humano; Elena de Miguel y Olimpia Pérez, antiguas alumnas y ahora colegas, que me han ayudado con las clases, la memoria y otras mil cosas más; también, mi más profundo agradecimiento a mis compañeros en la Facultad de Informática José Luis Sierra, Antonio Navarro, Antonio Sarasa, Carmen Fernández, Luis Hernández que me han invitado a formar parte de sus equipos y proyectos de investigación, que han impulsado y enriquecido mi investigación, y me han proporcionado apoyo financiero; mis compañeros de la oficina del Campus Virtual: Jesús Cristóbal, Jorge Merino, Miguel Peralta, Yoli Roldán, siempre dispuestos a ayudarme con las múltiples dudas y problemas técnicos y con los tuve la oportunidad de trabajar para crear el Campus Virtual de la UCM bajo la dirección de Alfredo Fernández-Valmayor; David Carabantes, a quien agradezco que me haya ayudado a salvar los obstáculos del final de esta carrera; Sara Olmos, artista y diseñadora gráfica, a quien doy las gracias por el precioso diseño de la portada de la tesis y por el regalo de su disponibilidad; mis compañeros del Comité de Coordinación del Campus Virtual, con los que he compartido años de experiencia y trabajo en la enseñanza y aprendizaje virtual.

Quiero expresar mi sincero agradecimiento al Decano de la Facultad de Filología, Dámaso López, y a mis compañeros del equipo decanal, que han apoyado firmemente mi dedicación a la tesis; al personal del equipo de Informática y Teconologías de la Facultad, que trabajan o han trabajado conmigo estos últimos años y que me han cubierto en muchos momentos para que pudiera tener "huecos" de trabajo dedicado a la tesis. También, mi agradecimiento

a Amelia Sanz, compañera incansable en el Vicedecanato de Tecnologías; a Jorge Arús, compañero actual que con gran paciencia me ha sobrellevado en estos últimos meses de tesis; a mis compañeros de titulación, de Departamento, colegas y personal de la Facultad que me han ayudado a quitar algunas de las piedras que han aparecido en el camino. Son muchas las personas que, de una u otra forma, me han brindado su apoyo y que han facilitado con su profesionalidad y amabilidad mi trabajo de investigación.

Esta investigación se ha llevado a cabo dentro de los proyectos de investigación que han enmarcado y financiado mi trabajo investigador: "Objetos de aprendizaje en el Campus Virtual (OdA-Virtual)" (TIN2005-08788-C04-01)[1]; "Un modelo hipermedia modular para la enseñanza de la Lingüística General" (TIN2005-08788-C04-03)[2]; "Arquitecturas Avanzadas en Campus Virtuales (AACV)" (TIN2009-14317-C03-01/TSI)[3]; "Integración de plataformas y servicios en el campus virtual (IPS-CV)" (TIN2008-06708-C03-01/TSI)[4]; "Tecnologías de Marcado Descriptivo –XML- como base a un Proceso de Desarrollo de Software Guiado por Lenguajes" (UCM-Santander Central Hispano)[5]; y "Glosario interactivo para el aprendizaje de conocimientos jurídicos en el campus virtual abierto" (PIMCD 66/2008)[6].

He podido contar, además, con el apoyo de mis amigos y familia; mis amigos que, sin perder la paciencia, han seguido mi trabajo, me han ayudado con los niños e, incluso, se han encargado de organizar las pocas veladas que hemos podido pasar juntos; Gloria ha sido mi apoyo en casa durante todas estas tardes de trabajo…

Finalmente, quiero dar las gracias a mi familia, a mis padres, mis padres políticos; mis abuelos (que ya no están), mis hermanos, especialmente Enrique quien, con su saber y experiencia en bases de datos, me ha aportado valiosísimas ideas. Ellos me han soportado con admirable paciencia durante estos años de tesis, se han preocupado y ocupado de mis

[1] Duración:1/Enero/2006 hasta 1/Junio/2009. Investigador responsable: Afredo Fernández-Valmayor.
[2] Investigadora responsable: Covadonga López Alonso.
[3] Entidades participantes: Universidad Complutense de Madrid (Facultad de Informática), Universidad Nacional de Educación a Distancia (Facultad de Informática). Duración, desde: 01/Enero/10 a: 31/Diciembre/12. Investigador responsable: Antonio Navarro Martín.
[4] Entidades participantes: Universidad Complutense de Madrid (Facultad de Informática). Duración: 01/Enero/09 a: 31/Diciembre/09. Investigador responsable: Antonio Navarro Martín.
[5] Entidades participantes: Facultad de Filología de la UCM, Facultad de Informática UCM, Duración desde: 1/Enero/2008 hasta: 31/Diciembre/2009. Investigador responsable: José Luis Sierra Rodríguez.
[6] Entidades participantes: Facultad de Filología; Facultad de Derecho; Facultad de Ciencias Físicas de la UCM, Duración desde 1/2/2008 a 31/12/2008.

tareas familiares, regalándome todo el tiempo que he necesitado (y ha sido mucho) para hacer este trabajo; a mi marido, Miguel, y mis hijos, Miguel y Javier, les doy las gracias, además, por haber renunciado a tantas horas juntos, por haberlo comprendido e, incluso, haber hecho de este trabajo de tesis un proyecto familiar más.

A todos, mi más sincero agradecimiento.

Índice General

Resumen

Este trabajo puede catalogarse como una contribución dentro de la Lingüística Computacional a la Tecnología Educativa, concretamente al *e-learning*. El objetivo es facilitar la construcción de los tesauros académicos de explotación en formato electrónico y, por ello, estos tesauros hay que entenderlos como sistemas lingüísticos para expresar y organizar el conocimiento de un dominio. En ellos se utilizan términos y relaciones semánticas del mismo lenguaje específico usado en los materiales o las colecciones de recursos docentes o de investigación creados por y para la actividad académica, siempre en entornos electrónicos de enseñanza y aprendizaje. El propósito de estos tesauros es: i) ayudar al profesor a organizar conceptualmente sus materiales, haciendo más fácil su localización, selección, y uso; y ii) ayudar al alumno a entender y aprender los conceptos y a usar de forma adecuada la lengua de especialidad de la disciplina o área de conocimiento que cubra el tesauro.

Nuestra propuesta es un modelo dinámico formal que representa, mediante estructuras relacionales, el contenido de los tesauros. Con él se da soporte a un método de construcción incremental e inductivo que genera los tesauros como parte del proceso de creación de materiales didácticos o de investigación, reproduciendo el modo en que los autores organizan y describen estos materiales. El modelo y método Higraph Léxico proporcionan el fundamento para la creación de aplicaciones informáticas de carácter general que sirvan para que los profesores, investigadores y estudiantes puedan crear, visualizar, manipular y actualizar automáticamente sus tesauros académicos de explotación.

Abstract

This work can be considered as a contribution, within Computational Linguistics, to Educational Technology, specifically to e-learning. The aim is to facilitate the construction of academic thesauri for electronic exploitation. These thesauri are, therefore, to be understood as linguistic systems for the expression and organization of a domain's knowledge. They use the same terms and semantic relations as the language found in materials or series of teaching and research resources created by and for academic activity, always within the realm of electronic teaching and learning. The aim of these thesauri is to i) help the teacher organize didactic – as well as research – materials conceptually, thus facilitating the localization, selection and use thereof; and ii) help the student understand and learn concepts and use, accurately, the language specific to the discipline or field of knowledge covered by the thesaurus.

Our proposal is a dynamic formal model which represents, by means of relational structures, the highly intertwined and changeable contents of thesauri. It gives support to an incremental and inductive construction method which generates thesauri as part of the creation process of teaching and research materials and which replicates the way authors organize and describe those materials. The HL model and method provides the foundation for the creation of general computer applications which may help teachers, researchers and students automatically build, visualize and update their thesauri for academic exploitation.

Capítulo 1

Marco de la tesis

1.1 Panorama general

El presente trabajo se ha desarrollado dentro de un Proyecto de Investigación y Desarrollo Tecnológico financiado por el Ministerio de Educación y Ciencia con título "Objetos de aprendizaje en el Campus Virtual (OdA-Virtual)" (TIN2005-08788-C04-01)[1] en el que han participado las Facultades de Informática, Filología, y Geografía e Historia. El objetivo del proyecto, recientemente terminado, era desarrollar los procesos, metodologías, plataformas y arquitecturas que den soporte a la participación de profesores y alumnos en la construcción y utilización de Objetos de Aprendizaje (OdA) en el entorno de un campus virtual. Ya desde el inicio del proyecto se había detectado la necesidad de disponer de un nuevo tipo de tesauros académicos de explotación para la clasificación, indexación y búsqueda de los OdA que los profesores creaban y utilizaban de forma colaborativa en los entornos de enseñanza y aprendizaje virtuales. Esta necesidad motivó este trabajo de tesis. Los trabajos de investigación interdisciplinares (Informática, Lingüística e Historia) de los tres subproyectos dieron el soporte y marco de aplicación necesario para el planteamiento y desarrollo de esta tesis:

1) el subproyecto de Informática tenía como objetivo desarrollar la base teórica sobre la que fundamentar la definición y construcción de los procesos, metodologías y plataformas que deben constituir el entorno que permita a profesores y alumnos elaborar OdA;
2) el subproyecto de lingüística, "Un modelo hipermedia modular para la enseñanza de la Lingüística General" (TIN2005-08788-C04-03) [2], tenía como objetivo básico el desarrollo de una metodología modular que permita analizar y estructurar la información contenida en los OdA en el marco de un campus virtual, para apoyar la enseñanza y la investigación; y
3) el subproyecto de Geografía e Historia, "Estudio y construcción de Objetos Virtuales en Geografía e Historia" (TIN2005-08788-C04-04)[3] tenía como objetivo estudiar la construcción de OdA que se puedan componer y que

[1] Duración:1/Enero/2006 hasta 1/Junio/2009. Investigador responsable: Afredo Fernández-Valmayor.
[2] Investigadora responsable: Covadonga López Alonso.
[3] Investigadora responsable: Mercedes Guinea Bueno.

integren los trabajos de investigación y el material docente generado por los profesores de esta área de conocimiento dentro del marco del Campus Virtual de la UCM. Con este fin, este grupo llevó a cabo una línea de investigación basada en la reutilización del material gráfico, documental y museístico existente en el Departamento de Historia de América II (museo, laboratorio, archivos, informes de investigación y/o notas de clase) y en los otros centros de investigación participantes (CNRS y University of Texas en San Antonio) para la realización y distribución en la web de estos OdA.

Posteriormente, otros proyectos han permitido, y están permitiendo, aplicar los resultados obtenidos en esta investigación, aportando una experiencia valiosa para corregir y mejorar de forma incremental la propuesta inicial: a) "Arquitecturas Avanzadas en Campus Virtuales (AACV)", financiado por el Ministerio de Ciencia y Tecnología (TIN2009-14317-C03-01/TSI)[4]; b) "Integración de plataformas y servicios en el campus virtual (IPS-CV)", financiado por el Ministerio de Ciencia y Tecnología (TIN2008-06708-C03-01/TSI)[5]; c) "Tecnologías de Marcado Descriptivo –XML- como base a un Proceso de Desarrollo de Software Guiado por Lenguajes", financiado por UCM-Santander Central Hispano[6]; y d) "Glosario interactivo para el aprendizaje de conocimientos jurídicos en el campus virtual abierto" (PIMCD 66/2008), financiado por el Vicerrectorado de Desarrollo y Calidad de la Docencia de la UCM[7].

A todos ellos nuestro agradecimiento.

1.2. Introducción

En la actividad universitaria surge la necesidad de expresar y organizar el conocimiento y las creaciones intelectuales desarrolladas o difundidas por los profesores, investigadores y estudiantes en entornos electrónicos de formación universitaria, los campus virtuales, con este tipo de repertorios que denominaremos *tesauros académicos de explotación*.

---

[4] Entidades participantes: Universidad Complutense de Madrid (Facultad de Informática), Universidad Nacional de Educación a Distancia (Facultad de Informática). Duración, desde: 01/Enero/10 a: 31/Diciembre/12. Investigador responsable: Antonio Navarro Martín.

[5] Entidades participantes: Universidad Complutense de Madrid (Facultad de Informática). Duración: 01/Enero/09 a: 31/Diciembre/09. Investigador responsable: Antonio Navarro Martín.

[6] Entidades participantes: Facultad de Filología de la UCM, Facultad de Informática UCM, Duración desde: 1/Enero/2008 hasta: 31/Diciembre/2009. Investigador responsable: José Luis Sierra Rodríguez.

[7] Entidades participantes: Facultad de Filología de la UCM (Area de Lingüística General), Facultad de Derecho, Facultad de Ciencias Físicas. Duración desde: 1/Febrero/2008 hasta: 31/Diciembre/2008.Investigador responsable: María de la Sierra Flores Doña. Coordinadora del subproyecto de informatización: Ana Fernández-Pampillón.

Un *tesauro* es un vocabulario limitado, generalmente de palabras especializadas, dotado de sus correspondencias semánticas, y elegido para que represente las nociones que figuran en un texto dado para su empleo en informática y en el establecimiento de índices (Martínez de Sousa, 1995). Los términos de un tesauro están formalmente organizados de forma que se hacen *explícitas* las relaciones entre los conceptos, por ejemplo, de hiponimia-hiperonimia. Las relaciones estándar entre los términos de un tesauro son las relaciones semánticas de equivalencia[8], jerárquicas y asociativas, y se visualizan mediante marcadores estándares y recíprocos (ANSI/NISO Z39.19, 2005).

Los tesauros son herramientas lingüísticas que sirven para ayudar a las personas o a las máquinas a encontrar los términos más apropiados para expresar una idea (Aitchinson y Clarke, 2004). Sus aplicaciones más frecuentes son de tipo (i) lingüístico, (ii) documentalista, (iii) informático y (iv) académico.

i) Desde el punto de vista lingüístico, el tesauro se concibe como una herramienta de soporte para el escritor que le ayuda a encontrar los términos más adecuados a la idea que quiere expresar en sus composiciones literarias. Un ejemplo paradigmático es el tesauro de Roget[9], actualmente disponible en formato papel y electrónico[10]. El tesauro de Roget no organiza los términos alfabéticamente como en otros vocabularios tradicionales, diccionarios, glosarios, enciclopedias, sino que se agrupan de forma sistemática según los conceptos que expresan. De esta forma el usuario puede ir "desde la idea a la palabra; desde la palabra a la idea" (Casares, 1959).

   La estrategia de búsqueda en cualquier tesauro es similar a la de un diccionario ideológico[11] cuando se quiere encontrar el término más adecuado a una idea: primero, debe expresarse con toda claridad el problema, o la cuestión, cuya solución interesa buscar con algún término o términos; segundo, si la organización sistemática del tesauro –o del diccionario ideológico- es correcta a partir del término o términos que expresan el problema o bien se encuentra la

[8] En los estándares de tesauros para la Recuperación de Información no se incluyen relaciones de oposición ni de homografía.

[9] *Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition*, de Peter Mark Roget publicado en 1852.

[10] Se puede consultar una versión en línea en: http://poets.notredame.ac.jp/Roget/contents.html

[11] En la lexicografía española, el diccionario ideológico es un tipo de diccionario onomasiológico -parte de los conceptos-, cuyos lemas están ordenados alfabéticamente y encabezan un grupo de palabras que corresponden a un campo léxico determinado (Haensch, 1997: 67-68). En lengua inglesa, se entiende por diccionario ideológico un diccionario clasificado por temas. En lexicografía francesa, el diccionario ideológico se corresponde con el diccionario analógico.

serie de términos asociados a la consulta o bien se irá explorando la red del tesauro guiándose por las clasificaciones y las relaciones semánticas que conectan los términos. Las series de términos pueden incluso cruzarse cuando comparten sus significados. Para precisar el significado, los términos pueden incluir notas de ámbito[12] y también cualificadores[13] porque en un tesauro es primordial que cada término represente un solo concepto. La búsqueda puede hacerse en profundidad, siguiendo las relaciones de tipo hiponimia-hiperonimia, o en anchura, siguiendo las relaciones de tipo asociativo.

En el caso inverso, cuando se quiere encontrar el significado de una palabra en el tesauro, la forma de proceder es diferente de la habitual en otras obras lexicográficas que incluyen definiciones de los términos. En los tesauros el significado de un término viene determinado por las relaciones con los otros términos y por las notas de ámbito, si existen. "El tesauro presenta, a través de su estructura, una serie de relaciones que establece en general el contexto de "significado" de un término dado, con especial referencia a términos de connotación más amplia o más restringida. Normalmente esto basta para indicar la interpretación que se hace de un término. A veces, cuando un término se interpreta vagamente en el uso común o cuando diferentes diccionarios le asignan significados diversos, es necesario extender la nota de aplicación hasta que constituya una definición completa" (UNE 50106, 1990). El procedimiento consiste, en este caso, por lo tanto, en explorar el tesauro a partir de la palabra buscada: en primer lugar, se consultan los términos sinónimos y las notas de ámbito, si existen; en segundo lugar, los hiperónimos e hipónimos cercanos; y en tercero, los términos asociados.

ii) En el contexto de trabajo del documentalista, el tesauro es un *lenguaje documental*. Un lenguaje documental es un lenguaje controlado que se usa para representar la información contenida en un conjunto de documentos, con el fin de facilitar su almacenamiento y su posterior recuperación. El lenguaje se controla con reglas que normalizan la forma de los términos en tres niveles: en el nivel morfológico, definiendo la categoría gramatical o la flexión; en el nivel

[12] Es una explicación o definición de un término. Cuando el tesauro se utiliza para la indexación de documentos o de cualquier objeto con contenido informacional, las notas de ámbito sirven para indicar el uso que se le quiere dar en ese lenguaje de indexación.

[13]Un cualificador es otro término situado entre paréntesis que se añade al término para desambiguar su significado, indicando, por ejemplo, el área temática respecto de la cual debe interpretarse.

sintáctico, definiendo las combinaciones de términos, llamada coordinación; y en el nivel semántico, restringiendo el significado de los términos con las notas de ámbito y los cualificadotes para que un término sólo represente un concepto y un concepto sólo sea representado por un término. El objetivo es describir de forma precisa el contenido de los documentos, evitando la ambigüedad del lenguaje natural. El tesauro es un tipo de lenguaje documental *postcoordinado*, que permite la identificación de los documentos utilizando cualquier combinación de términos del tesauro, no necesariamente prefijada –como ocurre con los lenguajes *precoordinados*- con el fin de representar de forma flexible los documentos mediante la yuxtaposición de conceptos. Aquí radica uno de los puntos fuertes de los tesauros, respecto de otros tipos de lenguajes documentales precoordinados como las clasificaciones, proporcionan una descripción flexible y exhaustiva del contenido temático de los documentos que incrementa, con los múltiples puntos de vista que proporcionan las combinaciones de términos no prefijadas, las posibilidades de recuperación (Laguens, 2006).

iii) Este uso de los tesauros como lenguajes documentales se extiende también al contexto informático de la R*ecuperación de Información*[14]. La recuperación de información (RI) es una rama multidisciplinar[15] que estudia los mecanismos de representación, almacenamiento, organización y acceso a la información en colecciones de documentos, en los contenidos de los documentos, en bases de datos o en la Web (Baeza-Yates y Ribeiro-Neto, 1999). Los tesauros se incorporan en los sistemas de RI desde los años cincuenta con la idea de "transformar los conceptos y sus relaciones que se expresan en los documentos, en un lenguaje más regularizado, con los sinónimos controlados y las estructuras morfosintácticas simplificadas" (Brownson, 1957)[16]. De esta forma, la persona -o aplicación software- que indexa y la que busca utilizan un mismo lenguaje. Incluso cuando en la búsqueda de información se puede utilizar texto libre, el tesauro es útil para extender los términos de consulta con sinónimos o hiperónimos, de forma que aumenten las posibilidades de recuperar lo que se desea. Además, si se necesita restringir la búsqueda –para aumentar la precisión

[14] El tesauro fue utilizado por primera vez para la recuperación de información en 1956 por Peter Luhn de IBM (Aitchinson y Clarke, 2004).
[15] Algunas disciplinas involucradas son la Informática, la Psicología cognitiva, la Lingüística y la Biblioteconomía y la Documentación.
[16] Este texto es una de una de las primeras referencias al uso de los tesauros como herramientas de apoyo a la RI. Tomado de: (Aitchinson y Clarke, 2004; Gil, 1998a).

de los resultados- el tesauro proporciona los hipónimos de los términos de búsqueda. En definitiva, desde el punto de vista de la RI, el tesauro se utiliza como una herramienta de apoyo para indexar, clasificar, buscar, o seleccionar información (Lancaster, 1986).

La Web es, actualmente, el soporte universal para la producción, almacenamiento y difusión de la información. Este soporte, sin embargo, carece de mecanismos de carácter general para organizar y describir de forma coherente el gran volumen de información (Berners-Lee et al., 2001), por lo que es un problema recuperar, operar e integrar tanta información y tan heterogénea (Stuckenschmidt, van Harmelen, 2005)[17]. La aplicación de tesauros –y otros tipos de vocabularios[18] como categorizaciones, taxonomías y ontologías- para la identificación y organización de la información es un mecanismo que mejora la efectividad en la recuperación de la información en la Web, no sólo porque permite expandir los términos de consulta para lograr una mayor exhaustividad o precisión en las búsquedas, sino también porque proporciona una descripción del marco conceptual de la información en un sublenguaje controlado del lenguaje natural (Soergel, 2002). Combinado esto último con la posibilidad de crear estructuras hipertextuales en la Web, permite construir tesauros electrónicos –accesibles en línea- para visualizar un "mapa terminológico-conceptual" en el que navegar, explorar y seleccionar los contenidos de información que se desean (Aitchison et al., 2000). Ejemplos paradigmáticos de este uso son los tesauros que se incorporan en los motores de búsqueda Web de las Bibliotecas Digitales – por ejemplo el tesauro de la Biblioteca de la UCM[19]- o de grandes de bases de datos documentales –tesauro EUROVOC[20]- o en los motores de búsqueda Web en Internet –Simpli[21].

Otra de las aplicaciones de los tesauros, derivadas de su concepción como mapa terminológico-conceptual, es la de facilitar la combinación de información heterogénea en Internet (Stuckenschmidt, van Harmelen, 2005; Soergel, 2002). Los tesauros, y también las ontologías, son utilizados por las personas o las aplicaciones software como modelos formales compartidos de un dominio de

[17] Este aspecto se trata en el prefacio y el capítulo 1.
[18] Utilizamos el término vocabulario para referirnos a estos tipos de repertorios porque es el término utilizado en los estándares de construcción de tesauros.
[19] http://alfama.sim.ucm.es/tesauro/tesauroPublic.htm
[20] http://europa.eu/eurovoc/
[21] http://www.simpli.com/

información respecto del cual se refieren e interpretan las diferentes fuentes de información. Se trata, en este caso, de utilizar el tesauro para: (i) calcular el grado de cercanía semántica de los contenidos de información de las diferentes fuentes de información respecto de la consulta del usuario; (ii) proporcionar un lenguaje común de consulta para las múltiples fuentes de información; y (iii) establecer correspondencias entre los términos de descripción de los contenidos de las distintas fuentes. Este tipo de aplicación de los tesauros se utiliza en los *repositorios digitales*[22] *federados* de recursos, que son múltiples almacenes de contenidos y de recursos web interconectados e integrados de forma "transparente" al usuario (IMS Digital Repositories, 2003). Normalmente, la aplicación de interconexión utiliza un tesauro u ontología general de referencia para establecer las correspondencias semánticas entre los contenidos de los distintos repositorios que pueden, incluso, tener tesauros propios más específicos.

iv) Finalmente, otra de las aplicaciones destacadas de los tesauros es la académica (Soergel, 2002). El tesauro (i) guía al estudiante en la búsqueda y asimilación de la información como parte integral del proceso de resolución de problemas en el aprendizaje y en el trabajo intelectual; (ii) proporciona al profesor un marco conceptual coherente para clasificar sus contenidos didácticos digitalizados, facilitando su almacenamiento, recuperación y uso posterior en entornos electrónicos de enseñanza y aprendizaje (plataformas *e-learning*); (iii) ayuda al investigador en la formulación, exploración y estructuración del contexto conceptual de la cuestión o hipótesis de investigación[23], ya que proporciona clasificaciones consistentes de las distintas aproximaciones, variables o criterios sobre un tema y el estado de la cuestión.

1.3. Motivación

Lo que ha motivado este trabajo de tesis es la necesidad de definir, en los entornos académicos universitarios, una nueva forma de entender, construir y usar los tesauros. El tesauro, desde este punto de vista, es *un instrumento para sistematizar y expresar el*

[22] Un repositorio digital es una colección de recursos accesibles mediante una conexión en red en la que no es necesario conocer cuál es la estructura de la colección. En esto último se diferencia de las bases de datos, en las que es imprescindible conocer la estructura de las colecciones para acceder y gestionarlas.
[23] Por ejemplo, ayuda a definir las dimensiones de un problema y los aspectos que deben considerarse en su resolución.

conocimiento desarrollado o recopilado, individual o colectivamente, durante la investigación o el aprendizaje sobre un tema o una disciplina. Esta necesidad nace, probablemente, de la reciente disponibilidad de los entornos y herramientas TIC necesarios para que el profesor pueda, de forma eficaz, construir y poner en marcha sus propios recursos didácticos y de investigación. Para referirnos a esta nueva concepción del tesauro introducimos el término *tesauro académico*. Un tesauro académico, por lo tanto, se distingue de otros tesauros porque es un tesauro creado por profesores, investigadores y estudiantes con el conocimiento y lenguaje propios de una determinada área de especialidad con el fin de utilizarlo, principalmente, en un entorno académico.

Este aprovechamiento es variado: (i) para organizar conceptualmente los materiales didácticos y de investigación del profesor, haciendo más fácil su localización, selección y uso; y (ii) para la enseñanza de los conceptos y el manejo de la lengua de especialidad de la disciplina o área de conocimiento que cubra el tesauro[24]. En cualquiera de los casos, se trata de utilizar el tesauro para *explotar* los conocimientos, contenidos o recursos didácticos y de investigación y, por lo tanto, lo denominamos *tesauro académico de explotación*.

Actualmente, el conocimiento, los contenidos y los recursos educativos se difunden y utilizan en los campus virtuales, especialmente en el contexto académico universitario. Los *campus virtuales* son espacios electrónicos en Internet, creados con plataformas *e-learning*, donde los profesores y alumnos interaccionan para enseñar y aprender e incluso, investigar. Esta interacción se denomina *enseñanza y aprendizaje electrónico (e-learning)* y, por ello, en los campus virtuales, los contenidos y los recursos didácticos están digitalizados y el conocimiento se difunde "digitalmente". Normalmente, los responsables de la creación, almacenamiento, clasificación y uso del conocimiento, contenidos y recursos son los profesores.

Los *tesauros académicos para la explotación* de materiales y recursos didácticos en entornos digitales deben representar los contenidos de estos materiales y recursos utilizando el lenguaje específico de los profesores para que sean realmente útiles. Y aquí radica el problema: es muy difícil disponer de tesauros con un alcance y naturaleza ajustados a las necesidades del profesor. Además, la búsqueda del tesauro más

[24] Mediante actividades didácticas colaborativas que favorezcan la consulta y estudio del tesauro. Por ejemplo, las actividades de construcción, exploración y búsqueda de términos o de materiales didácticos permiten que el alumno se familiarice con los términos y las relaciones conceptuales propias de un dominio o especialidad.

apropiado, su estudio y el uso de *tesauros de referencia*[25] supone un esfuerzo grande sin garantías de que vaya a ser rentable: la experiencia indica que los usuarios tienen dificultades para comprender y aplicar estos recursos lingüísticos en la clasificación de sus contenidos y recursos digitalizados (CEN CWA 14871, 2003).

La falta de precisión en la definición del dominio -de conocimiento, contenidos o recursos- que se necesita explotar, y los desajustes entre el lenguaje del tesauro y el lenguaje de los usuarios, profesores, investigadores y estudiantes, restan efectividad a estos tesauros (Lancaster, 1986:157). Algunos de los problemas que surgen son:

(1) la dispersión de datos: en la colección aparecen constantemente palabras que el tesauro no es capaz de normalizar[26] (Pérez Agüera, 2004);

(2) la ambigüedad semántica es excesiva, incluso en tesauros de dominio específico (Pérez Agüera, 2006); y

(3) los desajustes conceptuales entre la estructura (categorías y relaciones semánticas) del tesauro y la concepción que tiene el usuario de ese dominio[27] (Gruninger y Lee, 2002).

Las soluciones posibles son, o bien adaptar los tesauros disponibles, o bien crear tesauros nuevos (Aitchinson et al., 2000). En ambos casos, se trata de un proceso complejo porque requiere amplios conocimientos en modelos y metodología de construcción de tesauros y en modelos y metodologías informáticas. La construcción y mantenimiento de tesauros son, además, procesos costosos, porque necesitan una prolongada y considerable inversión de tiempo y de recursos materiales y personales; en consecuencia, en pocos casos estas soluciones están al alcance de los profesores que, aunque son especialistas en su materia y en enseñar, no lo son en tesauros o en informática.

Además, los tesauros electrónicos requieren modelos y aplicaciones informáticas para su creación, mantenimiento y gestión. El dominio léxico, en general, es un dominio complejo que contiene una gran cantidad y tipología de relaciones y que está en permanente evolución, con cambios que afectan no sólo al contenido sino también a la estructura. Desde el punto de vista informático, la construcción de los tesauros

[25] Tesauro de libre acceso y uso, construido por un comité de expertos oficialmente constituido para que sirva de referencia en el dominio o especialidad, con el objetivo de unificar el lenguaje y favorecer la interoperabilidad.

[26] No es posible resolverlo con una actualización periódica hecha a mano en función del crecimiento de la colección.

[27] Los tesauros de referencia constituyen una conceptualización elaborada y consensuada por un grupo reconocido de especialistas [ANSI/NISO Z39.19, 2005] que normalmente no son los usuarios finales de dicho vocabulario.

electrónicos actuales presenta dos problemas que afectan a la efectividad del tesauro y que limitan la disponibilidad de herramientas software de carácter general:

1) los modelos de datos con capacidad de expresar de forma completa estructuras de información complejas en permanente cambio como los modelos basados en grafos no son modelos suficientemente eficientes y, viceversa, los modelos de datos más eficientes, por ejemplo, el modelo relacional, son modelos con menos capacidad de representación conceptual.

2) los modelos de datos no son suficientemente generales como para obtener esquemas de datos uniformes e independientes del dominio que permitan un tratamiento uniforme del tesauro. Los tesauros se diseñan mediante técnicas de análisis y clasificación aplicadas al dominio de conocimiento -métodos deductivos- o al conjunto de términos fuente -métodos inductivos. El resultado es la producción de esquemas de organización ajustados al contenido *previsto* del tesauro. Estos esquemas de organización se traducen a esquemas de datos informáticos, aplicando algún modelo de datos adecuado para ese esquema y para los objetivos del tesauro. Los sistemas informáticos para construir y gestionar los tesauros necesitan utilizar estos esquemas de datos fijos para poder interpretar correctamente el contenido del tesauro. Pero las continuas modificaciones que surgen en el ámbito del tesauro no sólo cambian el contenido sino que también pueden afectar a la estructura de datos prevista inicialmente. Modificar el esquema de datos puede suponer rehacer todo el tesauro, porque los datos organizados con un esquema antiguo pueden no ser coherentes con un esquema de organización nuevo. En consecuencia, las posibilidades de construcción, actualización, intercambio y reutilización de los tesauros están limitadas por el uso de un esquema de datos inicialmente establecido.

1.4. Objetivos e hipótesis de trabajo

1.4.1. Objetivos

El objetivo de este trabajo es definir una nueva forma de entender y construir los nuevos *tesauros académicos de explotación*, *tesauros de especialidad, creados en formato electrónico por los profesores e investigadores, especialistas en su disciplina, con fines de explotación en actividades didácticas e-learning y/o actividades investigadoras*. Para ello es necesario encontrar un mecanismo, fácil de aplicar, para construir tesauros que sistematicen y expresen las ideas propias desarrolladas o recopiladas en contenidos o

recursos digitales, individual o colectivamente, durante la investigación, la enseñanza o el aprendizaje sobre un tema o una disciplina.

Este objetivo general se puede desglosar en los siguientes objetivos específicos:

1.- Encontrar estructuras del lenguaje de especialidad, las *estructuras terminológicas en semántica libre*[28], de forma abreviada, *estructuras-t*, que utilizan los profesores o autores para expresar las ideas que representan un dominio de conocimiento, de contenidos o de una colección de recursos. Por estructuras terminológicas en semántica libre nos referimos a pequeñas redes de términos con relaciones semánticas –una o varias simultáneamente- que no están previamente establecidas, que están inmersas en el contenido y/o meta-contenido de materiales educativos, y que son propuestas por uno o varios especialistas de esa comunidad de forma libre -por medio de una elección libre-, lo que no implica que sean originales o únicas.

2.- Buscar un modelo de datos informático general y flexible que sirva para recoger las estructuras-t en un sistema de signos formado por términos y categorías que están relacionados semánticamente y que está en permanente cambio. Este modelo podría también considerarse un meta-modelo para los tesauros, puesto que sirve para crear los esquemas conceptuales, ajustados al dominio, que estructuran los tesauros.

3.- Ofrecer una metodología que, utilizando el modelo anterior, sea capaz de construir sistemáticamente el tesauro a partir de las estructuras terminológicas, en semántica libre, de los contenidos o recursos digitales.

Este modelo y metodología deben servir de base para construir aplicaciones informáticas que, de forma general, puedan utilizar los equipos docentes para crear y gestionar sus tesauros académicos de explotación.

1.4.1. Hipótesis de trabajo

Para la consecución de estos objetivos se plantea las siguientes hipótesis de trabajo:

Si se considera que:

1) la lengua es un sistema estructurado de signos en el que el *valor* del significado de cada elemento depende de su posición diferencial respecto de los demás[29];

[28] Elegimos esta denominación por analogía con 'sintaxis libre' que supone estructuras sintácticas no consolidadas en la lengua como formas de cita (Lyons, 1977 pp. 22-26).

[29] Esta concepción sistémica del tesauro se basa en una semántica diferencial que tiene su origen en la propuesta de F. Saussure, "El valor de una palabra en su parte conceptual está constituida únicamente por sus conexiones y diferencias con los otros términos de la lengua […]" (Saussure, 1916: 220).

2) los tesauros son representaciones parciales de una lengua restringidos a las nociones de un dominio de conocimiento mediante términos organizados en grupos por relaciones semánticas; y
3) existe un modelo formal capaz de representar esta concepción de la lengua y del tesauro; en consecuencia
4) es posible representar de forma general y uniforme cualquier tesauro, con independencia de su naturaleza y aplicación, y es posible sistematizar el proceso de construcción y actualización de tesauros a partir de grupos de términos organizados por relaciones semánticas como las estructuras-t.

Teniendo en cuenta estos presupuestos, este trabajo puede catalogarse como una contribución dentro de la Lingüística Computacional a la Tecnología Educativa, concretamente al *e-learning*, cuyo objetivo es facilitar la construcción de los tesauros electrónicos, entendidos como sistemas lingüísticos de representación del contenido de un dominio, utilizando el mismo lenguaje específico con el que se expresa el conocimiento sobre los materiales o las colecciones de recursos docentes o de investigación creados por y para la actividad académica en los entornos electrónico de enseñanza y aprendizaje.

1.5. Metodología de trabajo

La metodología aplicada para la consecución de los objetivos y la demostración de la hipótesis consta de las seis etapas siguientes:

1. establecimiento de la cuestión de investigación;
2. análisis del estado del arte:
   2.1. análisis de los vocabularios y los tesauros desde el punto de vista lingüístico, documentalista e informático, en particular los vocabularios y tesauros de explotación;
   2.2. análisis de los entornos académicos de trabajo *e-learning*, en particular cómo se crean y utilizan los contenidos didácticos o de investigación y las colecciones de recursos educativos;
   2.3. análisis de las características y requisitos de los tesauros de explotación;
   2.4. análisis de los modelos estándares de construcción de tesauros monolingües;
   2.5. análisis de los modelos informáticos de representación de tesauros: los enfoques teóricos y sus aplicaciones al *e-learning*; y
   2.6. análisis de los métodos de construcción de tesauros de explotación;

3. planteamiento de la hipótesis de trabajo;
4. observación, recogida de datos y estudio de antecedentes utilizando fuentes de tipo:
   4.1. tecnológico-educativo: observación y estudio de los procesos de creación, clasificación y uso de contenidos y recursos didácticos en entornos virtuales (proyecto OdA). Experiencias directas en el Campus Virtual UCM. Revisión de otras experiencias en universidades y organismos; y
   4.2. lexicográfico y documentalista. Observación y estudio de los métodos de construcción y uso de vocabularios en general, y de vocabularios aplicados a la recuperación de información y la explotación académica en entornos digitales accesibles en la Web: campus virtual, bibliotecas digitales, repositorios de recursos educativos y bases de datos documentales;
5. método de demostración:
   5.1. modelo: planteamiento y desarrollo;
   5.2. modelo: experimentación[30];
   5.3. modelo: evaluación, ajustes y primeras conclusiones;
   5.4. método: planteamiento y desarrollo;
   5.5. método: experimentación[31]; y
   5.6. método: evaluación, ajustes del método y conclusiones del método
6. Estudio de resultados y establecimiento de las conclusiones finales

1.6. Estructura de la memoria

Hemos organizado la memoria en nueve capítulos. En este primer capítulo se establece el marco general de la tesis explicando el contexto de investigación donde se ha integrado esta tesis, una introducción sobre el objeto de estudio que son los tesauros, las cuestiones que han motivación de esta investigación, los objetivos e hipótesis del trabajo, la metodología aplicada en la investigación y, finalmente, la descripción de la estructura de esta memoria.

El segundo capítulo, los vocabularios para la explotación de recursos didácticos digitalizados, revisa el papel que juegan los vocabularios en la explotación de los

[30] Aplicación del modelo a una muestra de tesauros y vocabularios ya existentes, uno de referencia: el tesauro europeo ETB en su versión española y dos tesauros académicos de explotación: (i) el vocabulario del repositorio CHASQUI creado por el equipo de investigación de Geografía e Historia, y (ii) el tesauro del glosario explicativo sobre derecho electrónico creado por un equipo de profesores de la Facultad de Derecho.

[31] Aplicación del método a la misma muestra de tesauros que se utilizó para la experimentación del modelo.

recursos didácticos digitalizados; para ello, se revisa el concepto interdisciplinar de vocabulario, de vocabulario controlado, vocabulario de explotación y los tipos de vocabularios de explotación, entre los que se encuentran los tesauros de explotación de recursos didácticos en entornos *e-learning*.

En el tercer capítulo, los entornos virtuales de enseñanza y aprendizaje, se presenta una síntesis de los conceptos relacionados con el *e-learning* puesto que es el contexto donde surgen y se utilizan los vocabularios electrónicos, en general, y los tesauros académicos de explotación, en particular, como sistemas de referencia para la explotación del conocimiento creado por los profesores, investigadores y estudiantes en su actividad académica.

El cuarto capítulo, el modelo de los estándares de construcción de tesauros de explotación, lo dedicamos, fundamentalmente, a revisar este modelo que establece la naturaleza del contenido, los modos de presentación, y las reglas de modificación en los tesauros monolíngües; el capítulo se completa con una revisión sobre el concepto y los tipos de modelos de datos; las características y requisitos de los tesauros de explotación, y los modelos tradicionales alfabético y sistemático.

En el capítulo quinto, los modelos informáticos para la construcción de tesauros de explotación, se revisan los modelos de datos más utilizados para la construcción de tesauros y se analizan respecto a las características y requisitos de los tesauros de explotación.

El capítulo sexto, el modelo higraph léxico para la construcción de los tesauros, presenta nuestra propuesta de modelo general para la representación sistemática y visual del contenido de los tesauros; previamente, se introducen los modelos matemático y visual de los higraph y lingüístico del significado de los signos que constituyen el fundamento de la propuesta.

El capítulo séptimo, una metodología para la construcción inductiva de tesauros académicos de explotación, revisa, en primer lugar, los métodos generales de construcción de tesauros y, en segundo lugar, presenta la metodología nueva de construcción inductiva de los tesauros académicos de explotación que proponemos y que está basada en el modelo higraph léxico y en las estructuras-t creadas por los profesores.

En el capítulo octavo, casos prácticos, se presenta la experimentación del modelo y método propuestos en los capítulos anteriores con tres tipos de tesauros de explotación académica que son diferentes en propósito, tipos de estructuras-t y resultados.

El capítulo noveno recoge una recopilación de toda la investigación, las conclusiones finales y las líneas de trabajo futuro.

La bibliografía recoge las referencias en las que se ha basado el análisis del estado de la cuestión y las relativas a la línea de investigación. El apéndice A muestra la lista de tesauros utilizados en esta memoria, y el apéndice B el código del esquema de datos relacional del modelo HL que proponemos.

Capítulo 2

Los vocabularios para la explotación de recursos didácticos digitalizados

Desde la idea a la palabra; desde la palabra a la idea

(Casares, 1942)

Entendemos por explotación de recursos didácticos digitalizados el utilizarlos eficazmente, mediante la informática y las Tecnologías de la Información y Comunicaciones (TIC), para obtener el máximo provecho académico. Para ello es imprescindible que las personas y las aplicaciones informáticas sean capaces de acceder y entender fácilmente qué contienen estos recursos, que suelen estar almacenados en colecciones digitales poco accesibles por su gran tamaño. Este capítulo describe el papel que juegan los vocabularios en la explotación de los recursos didácticos digitalizados. Para ello, se revisa: 1) el concepto interdisciplinar de vocabulario, de vocabulario controlado, y su contenido; y 2) los tipos de vocabularios, entre ellos los tesauros, aplicados a la explotación de los recursos didácticos en entornos de enseñanza y aprendizaje electrónico, *e-learning*.

La primera cuestión, concepto y naturaleza de los vocabularios, se trata en las secciones segunda y tercera: la sección segunda, "Definiciones del término vocabulario", revisa el significado los términos vocabulario y vocabulario controlado, en las disciplinas de Lingüística y Tecnología Lingüística, Tecnología Educativa, Recuperación de Información y Biblioteconomía y la Documentación. La tercera sección, "El contenido semántico de los vocabularios", analiza los tipos de relaciones semánticas que pueden contener en los vocabularios.

La segunda cuestión, los tipos de vocabularios y su aplicación a la explotación e-learning, se trata en las tres secciones restantes de la forma siguiente: la cuarta sección, "Los vocabularios en los sistemas de recuperación de información", describe el papel que juegan los vocabularios en los sistemas de RI, especialmente en los procesos de indexación, búsqueda y navegación. La quinta sección, "El uso de vocabularios para la explotación didáctica de recursos digitalizados", revisa las aproximaciones actuales a la representación semántica de los recursos digitalizados usando metadatos y/o vocabularios En la sexta sección, "Tipos de vocabularios para la explotación de recursos didácticos digitalizados", se describen los tipos de vocabularios y su aplicación a la

recuperación de recursos educativos. Finalmente, en la séptima y última sección, se resume y se presentan algunas conclusiones del capítulo.

2.1 Introducción

Los vocabularios son recursos lingüísticos que permiten acceder al conocimiento a través de la palabra (Bougarev, 1996). Constituyen un mecanismo para organizar la información de un modo flexible y especialmente adecuado para entornos de trabajo en los que la información se crea de forma colaborativa y libre como en los Campus Virtuales (CV) universitarios centrados en el profesor. Sin embargo, es preciso tener en cuenta que el concepto de vocabulario es ambiguo, porque depende de la disciplina y de la aplicación. Para la construcción de un vocabulario es imprescindible la definición precisa de su naturaleza y objetivos. En caso contrario, se corre el riesgo de que los resultados sean un mero recopilatorio de palabras, no uniforme, incompleto y poco coherente que restan eficacia al vocabulario.

Un vocabulario o léxico[1] se define, desde el punto de vista lingüístico, como "… (1) el conjunto de palabras de un idioma; (2) un diccionario (libro); (3) el conjunto de palabras pertenecientes al uso de una región, de una profesión u oficio, de un campo semántico de un escritor, etc.", o simplemente, (4) el libro en que se contienen; …" (DRAE, 2001). Se trata de un término con un significado poco preciso y con un amplio contexto de aplicación[2]. Pueden distinguirse, además, varios tipos de vocabularios: (i) las listas de términos, (ii) los glosarios, (iii) las clasificaciones y taxonomías, (iv) los tesauros, (v) las ontologías, (vi) los diccionarios y (vii) los lexicones[3] (CEN CWA14871, 2003).

Cuando este inventario de palabras se sistematiza y administra adecuadamente el vocabulario sirve de herramienta para identificar, describir, acceder y explorar todos los *objetos digitales con un contenido* (documentos, sitios web, software, …) relativo a un dominio de conocimiento (Aitchison et al., 2000; Rodríguez y Ronda, 2005).

Sin embargo, la ambigüedad y la polisemia del lenguaje natural hacen inevitable la existencia de varios vocabularios para describir un mismo conjunto de objetos, con los consiguientes problemas de compatibilidad (Buckland et al., 1999). Los términos

[1] En el *DRAE 2001*, léxico es sinónimo de vocabulario en su tercera acepción.

[2] Fundamentalmente, en el procesamiento del lenguaje natural (Gibbon, 2000), clasificación conceptual (Garshol, 2004), clasificación documental (Buckland et al., 1999), indexación y recuperación de información (Lancaster, 1986).

[3] Lexicón se define como diccionario (DRAE,2001) y como léxico de una lengua (Martínez de Sousa,1995).

utilizados por los autores para describir el contenido de sus objetos digitales pueden no coincidir con los que se utilizan para organizarlos en los sistemas de almacenamiento y, probablemente, no coincidirán con los que utilizan los usuarios cuando los buscan. Esto último significa que los usuarios, en sus consultas, tienen que utilizar los mismos términos empleados por los autores e indexadores[4] para encontrar los objetos; para ello, o bien conocen el vocabulario de indexación, o bien tienen la capacidad de descubrir las varias y diversas formas de expresar un concepto. Entonces, ¿puede una persona expresar su petición con sus propias palabras y obtener el material que desea?

Los *vocabularios controlados* intentan recoger de las lenguas los términos que expresan cada concepto, seleccionar el más apropiado como preferido y realizar reenvíos desde los otros para conducir al usuario hasta el preferido. Cuando el vocabulario controlado se utiliza para la recuperación de objetos, éstos se indexan con los términos preferidos. De esta forma el usuario tiene libertad para buscar con cualquiera de los términos, preferidos o no preferidos. El vocabulario conducirá manualmente o automáticamente de la consulta a los objetos indexados. El vocabulario será útil sólo si sirve como lenguaje común de interfaz entre los términos de descripción de los objetos de contenido y los usuarios que buscan dichos objetos.

Además, algunos tipos de vocabularios controlados, como las taxonomías y los tesauros, agrupan los términos en categorías temáticas detalladas añadiendo una funcionalidad más a sus posibles aplicaciones: la clasificación u ordenación de los conceptos u objetos del dominio temático. Algunos autores consideran que un vocabulario controlado constituye un "mapa conceptual[5]" o un "esquema conceptual[6]" del dominio, que se puede utilizar como ayuda al usuario para sintetizar y relacionar los conceptos u objetos del ámbito del vocabulario e incluso como herramienta de exploración del dominio de conocimiento (Duncan, 1990; Jones et al., 1995; Garshol, 2004; Marzal et al., 2006).

[4] La indexación es el proceso de escoger los términos del vocabulario controlado que mejor describen los objetos de contenido, términos preferidos, y asociarlos con dichos objetos.

[5] Un mapa conceptual es una herramienta para representar y organizar gráficamente el conocimiento. Incluye conceptos y relaciones. Los conceptos se representan mediante etiquetas que, normalmente, son palabras o grupos de palabras (Novak y Cañas, 2008).

[6] Un esquema conceptual es la representación de una base de datos conceptual. Una base de datos conceptual es una abstracción del mundo real. Los esquema conceptuales incluyen tipos de entidades que se representan mediante etiquetas que, normalmente, son palabras o grupos de palabras, y tipos de relaciones entre los tipos de entidades (Ullman, 1988).

2.2. Definiciones del término vocabulario

Desde el punto de vista lingüístico y lexicográfico, la terminología utilizada para definir las distintas obras lexicográficas es, en general, poco precisa. Como se ha adelantado, resulta frecuente encontrar bajo un mismo término obras lexicográficas muy diversas entre sí (ver p. ej. Tablas 2.1, 2.2, 2.3). El término vocabulario, que ahora vamos a tratar, se utiliza con frecuencia como sinónimo de diccionario y de léxico y, al mismo tiempo, se usa tanto para referirse a obras que registran el léxico de una determinada área de conocimiento, materia, región, etc., como a inventarios de palabras ordenados alfabéticamente, lo que reflejan las definiciones de los diccionarios de uso del español más frecuentes[7] (Tabla 2.3).

Diccionario	Definición
(DRAE, 2001)	**1.** Libro en el que se recogen y explican de forma ordenada voces de una o más lenguas, de una ciencia o de una materia determinada. **2.** Catálogo numeroso de noticias importantes de un mismo género, ordenado alfabéticamente. *Diccionario bibliográfico, biográfico, geográfico.*
(CLAVE, 2002)	**1** Inventario en el que se recogen y definen las palabras de uno o más idiomas, generalmente por orden alfabético. ... SINÓN. *léxico* **2** Inventario en el que se recogen y explican los términos propios de una ciencia o de una materia, generalmente por orden alfabético. ...
(Seco et al., 1999)	**a)** Libro en que se recogen las palabras de una lengua, colocadas según un orden dado, gralm. alfabético, y acompañadas de su definición, explicación o equivalencia. **b)** *Con un compl especificador:* Libro en que se recogen las palabras (de una materia determinada), por orden alfabético y acompañadas de su definición, explicación o equivalencia.
(Moliner, 1998)	Libro en que se da una serie más o menos completa de las palabras de un idioma o de una materia determinada, definidas o con su equivalencia en otro idioma, generalmente por orden alfabético: 'Diccionario etimológico. Diccionario plurilingüe. Diccionario de sinónimos. Diccionario técnico'. ¤ Léxico, vocabulario. ¤ *Tratado de cierta materia en que los conceptos explicados están ordenados alfabéticamente: 'Diccionario de historia (o de filosofía)'
(Martínez de Sousa, 1995)	1) Recopilación de las palabras, locuciones, giros y sintagmas de una lengua o, dentro de ella, los términos de una ciencia, técnica, arte, especialidad, etc., generalmente dispuestos en orden alfabético (sin. abecedario, vocabulario). 2) Libro en el que al lado de las palabras de una lengua, generalmente colocadas en orden alfabético, figuran sus equivalentes en otras u otras lenguas.

[7] (DRAE, 2001), (CLAVE, 2002), (Seco et al., 1999), (Moliner, 1998), (Martínez de Sousa, 1995).

	3) Obra que ofrece por orden alfabético nombres, hechos, noticias, etc., referentes a un orden de conocimientos. 4) ABECEDARIO, cualquier lista cuyos términos aparecen en orden alfabético.

Tabla 2.1. Definición del término diccionario

Diccionario	Definición
(DRAE, 2001)	**1.** Catálogo de palabras oscuras o desusadas, con definición o explicación de cada una de ellas. **2.** Catálogo de palabras de una misma disciplina, de un mismo campo de estudio, etc., definidas o comentadas. **3.** Conjunto de glosas o comentarios, normalmente sobre textos de un mismo autor.
(CLAVE, 2002)	s. m. Catálogo de palabras oscuras, desusadas o técnicas, con definición o explicación de cada una de ellas. SEM. dist. de *léxico* (conjunto de palabras de una lengua; inventario de palabras de un idioma con definición).
(Seco et al., 1999)	*m* **1** Conjunto breve de palabras definidas o comentadas, pertenecientes a un texto o autor o a un ámbito determinado
(Moliner, 1998)	Catálogo de palabras, generalmente con una definición o explicación, sobre un asunto determinado, específicas de alguna disciplina, con alguna característica en común, etc. *Vocabulario.
(Martínez de Sousa, 1995)	1) Repertorio de voces cuyo fin es explicar un texto medieval o clásico, la obra de un autor, un texto dialectal, etc. 2) Repertorio no exhaustivo de palabras, generalmente técnicas, de una jerga determinada, como la ecología, la biología, la bibliología, etc.

Tabla 2.2. Definición del término glosario

Diccionario	Definición
(DRAE, 2001)	**1.** Conjunto de palabras de un idioma. **2. diccionario** (□ libro). **3.** Conjunto de palabras de un idioma pertenecientes al uso de una región, a una actividad determinada, a un campo semántico dado, etc. *Vocabulario andaluz, jurídico, técnico, de la caza, de la afectividad.* **4.** Libro en que se contienen. **5.** Catálogo o lista de palabras, ordenadas con arreglo a un sistema, y con definiciones o explicaciones sucintas.

	6. Conjunto de palabras que usa o conoce alguien. **7.** coloq. Persona que dice o interpreta la mente o dicho de otro. *Hablar por vocabulario. No necesitar de vocabulario.*
(CLAVE, 2002)	**1** Conjunto de palabras que componen una lengua o que pertenecen a una región, a una persona o a un campo determinados. ... SINÓN. *léxico* **2** Libro o lista en que se contiene este conjunto de palabras explicadas de una forma más o menos breve. ...
(Seco et al., 1999)	*m* **1** Conjunto de palabras (de un idioma). **b)** Conjunto de palabras propias (de una región, de una actividad, de un grupo humano o de una pers. determinados). **2** Catálogo ordenado y con definiciones sucintas de las palabras del vocabulario *esp* (1b).
(Moliner, 1998)	Serie de palabras reunidas según cierto criterio y ordenadas alfabética o sistemáticamente; por ejemplo, de palabras referentes a cierto oficio o de las precisas para redactar un tema o ejercicio en el aprendizaje de un idioma extranjero. Þ Tecnología, terminología. ¤ Serie alfabética de las palabras de una lengua. *Diccionario. ¤ Conjunto de palabras de una lengua. Léxico. ¤ Particularmente, el utilizado o conocido por una persona
(Martínez de Sousa, 1995)	1) Conjunto de palabras de un idioma. 2) Conjunto de palabras regionales, de una profesión u oficio, de un campo semántico, de un escritor, etc. 3) Libro en que se contienen los términos de un vocabulario. 4) Lista de palabras definidas sucintamente y colocadas por orden alfabético al final de un trabajo o un libro. 5) Diccionario

Tabla 2.3. Definición del término vocabulario

En otras disciplinas, en cambio, el significado y naturaleza de los vocabularios es más preciso y orientado a las aplicaciones, pero con diferencias entre ellas. La figura 2.1 muestra el contexto interdisciplinar en el que revisamos el concepto y uso de los vocabularios: la Lingüística, Biblioteconomía y Documentación, y las áreas tecnológicas de la Tecnología Lingüística (TL), la Tecnología Educativa (TE) y la Recuperación de Información (RI).

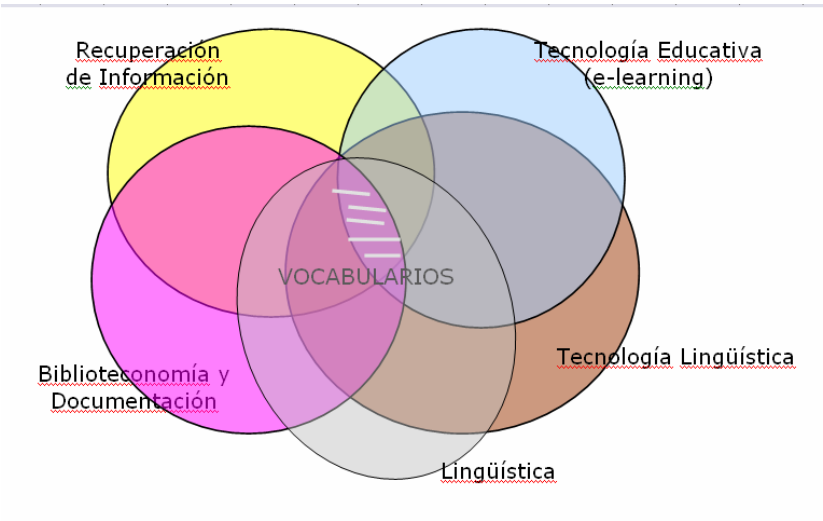


Figura 2.1. El contexto interdisciplinar de los vocabularios

En las áreas tecnológicas de *Recuperación de Información (RI), Tecnología Educativa (TE) y Tecnología Lingüística* (TL), los vocabularios se utilizan como componentes software que aportan una descripción conceptual y una dimensión pragmática y empírica del dominio de información.

Los Sistemas RI aplican, normalmente, *vocabularios controlados* para evitar la ambigüedad y polisemia del lenguaje (Lancaster, 1986). Un vocabulario controlado, como ya hemos mencionado, es una lista de términos enumerados explícitamente, no ambiguos y no redundantes. Esta lista es elaborada y mantenida por una "autoridad de registro" con los objetivos ideales de[8]:

1. *Traducir* cualquier término del lenguaje natural (utilizados por los autores, indexadores y usuarios) a los términos utilizados para indexar los objetos a recuperar.
2. Mantener la *consistencia* en los formatos y la asignación de términos.
3. Recoger y explotar las *relaciones semánticas* entre los términos.
4. Proporcionar un marco de *clasificación y navegación* que ayude a los usuarios a encontrar el objeto de contenido deseado. Y
5. Apoyar los procesos de *búsqueda y localización* de los objetos digitales con contenido.

Este concepto y uso de vocabulario procede, en realidad, del área de Biblioteconomía y Documentación. En esta disciplina los vocabularios son siempre controlados (Lewis y

[8] Ver especificación estándar de construcción de tesauros monolingües (ANSI/NISO Z39.19, 2005).

Sparck-Jones, 1996) y se definen como *lenguajes documentales*[9] que aportan un sistema común y universal de clasificación de las obras bibliográficas y de los documentos. Dentro de los vocabularios se distingue entre vocabularios precoordinados y postcoordiandos (Lancaster, 1986), como ya vimos, los vocabularios precoordinados están formados por términos y combinaciones de términos prefijadas para representar la materia o tema de cada documento del dominio. Normalmente se estructuran como vocabularios jerárquicos o asociativos. En esta categoría están los sistemas de clasificación y las listas tradicionales de materias (encabezamientos por materias) como el Sistema de Clasificación Decimal de Melvil Dewey, que fue creado en 1875 en Estados Unidos. Los lenguajes postcoordinados, por el contrario, están formados por términos y relaciones entre ellos que definen múltiples combinaciones posibles. Durante la fase de búsqueda se combinan los términos del vocabulario para obtener una combinación lo más cercana posible a la consulta del usuario. Los objetos digitales se indexan, por lo tanto, con tantos términos como se necesite. Los vocabularios usados en la RI y Biblioteconomía y Documentación tienden a ser, en la actualidad, postcoordinados, porque permiten una mayor libertad de consulta y menos "conocimiento" del lenguaje especializado por parte del usuario (Antelman et al., 2006)

Para la *Tecnología Lingüística* el vocabulario es un tipo de *recurso léxico*[10] que recoge, de una lengua (vocabularios monolingües) o varias lenguas (multilíngües), las palabras, sus relaciones, definiciones y otra información (Gibbon, 2000). Estos vocabularios pueden clasificarse en vocabularios en formato electrónico y vocabularios computacionales, lexicones computacionales. Los vocabularios en formato electrónico son digitalizaciones de los vocabularios en papel[11] que permiten capacidades de almacenamiento, prácticamente ilimitadas, y formas de acceso más rápidas y exactas a los contenidos. Sin embargo, las posibilidades de procesar automáticamente su

[9] Un lenguaje documental es un conjunto de términos o procedimientos sintácticos convencionales que se utilizan para representar el contenido de un documento con el fin de permitir su recuperación (Slype, 1991).

[10] "…El término recurso lingüístico se refiere a un conjunto de datos del habla o de las lenguas y sus descripciones en un formato legible para las máquinas, utilizado, por ejemplo, para la construcción mejora o evaluación de los sistemas o algoritmos de procesamiento del lenguaje natural y del habla o recursos para el software de búsqueda, para los estudios lingüísticos, la publicación electrónica, la traducción, etc. Ejemplos de recursos lingüísticos son los corpus de texto y habla, los lexicones computacionales, las bases de datos terminológicas,…" (ELRA, 2003).

[11] Esta forma es anterior a los vocabularios computacionales. Los primeros vocabularios electrónicos se crean en la década de los 80. Los vocabularios computacionales se comienzan a construir en los años 90, aunque los modelos y técnicas de construcción son muy anteriores, de los años 60. Los vocabularios computacionales utilizados hasta comienzos de los 90 eran demasidado pequeños (en media 36 palabras) como para ser considerados verdaderos vocabularios (Guthrie et al., 1996).

contenido son limitadas, ya que se reducen a operaciones de nivel morfosintáctico sobre las formas ortográficas[12], puesto que reproducen estructuras de organización del conocimiento léxico, previstas y preparadas para uso humano (Fernández-Pampillón y Matesanz, 2003)[13]. Se utilizan con fines primordialmente lingüísticos, aunque también se han aplicado, desde la tecnología lingüística, como fuente para extraer el conocimiento léxico para los vocabularios computacionales (Byrd et al., 1987; Walker et al., 1995)

Los vocabularios o lexicones computacionales, que son objeto de la Tecnología Lingüística y del Procesamiento del Lenguaje Natural (PLN)[14], se conciben como bases de datos y de conocimiento léxico diseñados para el procesamiento automático de las lenguas naturales (Allen, 1995). En estos vocabularios, el conocimiento léxico se hace explícito[15] y se organiza con modelos de datos informáticos que permiten un tratamiento automático más "inteligente", basado no sólo en operaciones a nivel morfológico y sintáctico sino también en la interpretación de los datos explícitos[16] (Brachman y Levesque, 1985; Bertino et al., 2001; Berners-Lee et al., 2001). Constituyen un componente básico en la arquitectura de los Sistemas PLN, y normalmente son accesibles para las personas a través de interfaces que abstraen las estructuras de los datos[17]. Son imprescindibles en el desarrollo de aplicaciones basadas en Tecnologías Lingüísticas como los correctores ortográficos y de estilo, la recuperación de información, el indexado y descripción de documentos y recursos (ELRA, 2003). Dos fuentes de distribución de lexicones computacionales son, por ejemplo, la agencia europea ELRA[18] y el consorcio americano LDC[19].

---

[12] Visualización de contenidos. Búsquedas basadas en la forma ortográfica: frases, sintagmas, grupos de palabras, palabras, prefijos y sufijos, caracteres y en el mejor de los casos variantes. Ordenación, importación exportación de contenidos, actualización del texto. Consulta de referencias marcadas explícitamente.

[13] El individuo es capaz de resolver las referencias implícitas, la falta de información y detectar las inconsistencias que contienen estas obras.

[14] Es una subárea de la Informática que tiene como fin la construcción de sistemas para el procesamiento automático de las lenguas naturales.

[15] El conocimiento léxico se clasifica tradicionalmente en siete niveles: conocimientos fonético y fonológico, morfológico, sintáctico, semántico, pragmático, de discurso y del mundo (Allen, 1995).

[16] En Brachman y Levesque: "…descripciones del dominio de modo que una máquina inteligente pueda llegar a conclusiones sobre su entorno manipulando formalmente estas descripciones…". La Lingüística computacional utiliza formalismos como las redes semánticas, estructuras de rasgos o predicados lógicos para representar formalmente el conocimiento léxico de forma explícita e implícita (resultados de las inferencias sobre estas estructuras explícitas) (Bertino et al., 2001; Allen, 1995; Shieber, 1986).

[17] WordNet es el ejemplo de lexicón computacional más referenciado [**wordnet**.princeton.edu].

[18] http://www.elra.info/

[19] http://www.ldc.upenn.edu/

Finalmente, en el área de la Tecnología Educativa los vocabularios son un mecanismo de representación semántica que permite a los agentes humanos o software localizar e interpretar contenidos educativos, bien para recuperar dicho contenido o bien para poder procesarlo con fines didácticos. Actualmente, los vocabularios se aplican, fundamentalmente, para resolver dos cuestiones: 1) la representación y recuperación de los recursos educativos digitalizados, y 2) la interoperabilidad entre herramientas y contenidos *e-learning* (Panizo et al., 2006; ANSI/NISO Z39.19, 2005; IMS-VDEX, 2004; CEN CWA 14871, 2003). En el primer caso, los vocabularios utilizados son del tipo taxonomías y tesauros (CEN CWA 14871, 2003), es decir, términos organizados en categorías y/o interconectados por relaciones semánticas de hiperonimia-hiponimia y otras. En el segundo caso, los vocabularios utilizados son del tipo tesauros y ontologías, para representar conceptualmente las dimensiones de un sistema *e-learning*: los agentes, las herramientas, el dominio de conocimiento, las metodologías y modelos de enseñanza, etc. (Sampson et al., 2004).

En este trabajo de investigación consideramos que el vocabulario es una herramienta de carácter lingüístico, informático y educativo que recoge y formaliza el conocimiento léxico de una lengua de especialidad, de forma que pueda ser utilizado por personas y sistemas informáticos, con el propósito de facilitar la comprensión y enseñanza de los términos y conceptos del área de especialidad, y la descripción y localización de recursos educativos digitalizados en la Web o en entornos e-learning[20] (Huynh, et.al., 2005).

2.3. El contenido semántico de los vocabularios

Un vocabulario, en cualquiera de sus interpretaciones, está formado como mínimo por una lista de términos[21], que pueden ser generales o específicos de un dominio (Hirst, 2004). Los términos que lo componen son palabras o grupos de palabras que pueden estar organizados en clases o categorías. La descripción de cada término está contenida en una *entrada léxica*. El contenido de las entradas depende del propósito del vocabulario (Gibbon, 2000), pero puede incluir, además de sus significados,

[20] En la web, los buscadores que incorporan los vocabularios como mecanismos de búsqueda son todavía escasos. Un ejemplo es el "Semantic Bank" en http://simile.mit.edu/bank/. Sin embargo, los repositorios de objetos de aprendizaje y los CMS ya incorporan vocabularios para la explotación de los contenidos que almacenan.

[21] En una o varias lenguas. En este trabajo se considerarán sólo vocabularios monolingües.

información gramatical, de uso, ortográfica, fonética, etimológica, relaciones con otros términos del vocabulario, etc.

En este sentido, un vocabulario puede concebirse también como un índice que hace corresponder la forma ortográfica de una palabra con la información sobre esa palabra (Hirst, 2004). No es, sin embargo, una correspondencia uno a uno, ya que en los casos en que una palabra tiene diferentes categorías sintácticas, en los homógrafos y en la polisemia, a una misma forma de palabra pueden corresponderle entradas diferentes. Los *vocabularios controlados* tratan de asegurar una correspondencia biunívoca mediante un proceso de control que desambigüa los términos homógrafos con un cualificador, de modo que selecciona, en caso de sinonimia, un único término como el término preferido o descriptor; y, en caso de términos polisémicos, restringe el significado mediante una nota de ámbito (ANSI/NISO Z39.19, 2005).

El contenido semántico de una entrada léxica se expresa en la definición. Puede tomar múltiples formas[22], desde un enunciado simple hasta reenvíos basados en el establecimiento de relaciones semánticas entre las unidades léxicas del vocabulario. Esta última aproximación, la definición del significado basado en relaciones semánticas, es preferida en los vocabularios orientados a la recuperación de información.

Las relaciones léxicas "clásicas" entre significados son la equivalencia, oposición, inclusión, co-hiponimia y parte-todo. *La relación de equivalencia* de significados es la *sinonimia*. Dos o más palabras son sinónimas si pueden sustituirse en cualquier contexto sin que cambie el significado (Lyons, 1977). Esta condición de igualdad en todos los significados de una palabra explica que los sinónimos absolutos sean escasos y que, a efectos prácticos, se prefiera una definición menos estricta, de modo que, dos o más palabras son sinónimas respecto de un significado si pueden intercambiarse en cualquier contexto[23].

La relación *opuesta* a la sinonimia es la *antonimia*. En la tradición lexicográfica, los antónimos son definidos como palabras de significados contrarios y, como tal, aparecen opuestos a los sinónimos[24]. Pero esta definición de antonimia es demasiado vaga. En Lehmann y Martin-Berthet, (1998) se afina la noción teniendo en cuenta que la antonimia implica una dimensión de parecido entre los términos. Los términos

[22] En Martínez de Sousa, 1995 se distinguen más de 40 tipos de definiciones.

[23] En Lyons, J., 1977 puede encontrarse una buena caracterización de la relación de sinonimia.

[24] Esta visión permite explicar la analogía de funcionamiento de los antónimos con los sinónimos: la sinonímia parcial y la antonímia parcial participan del mismo proceso ya que un término polisémico tiene, según sus acepciones y empleos, antónimos diferentes; por ejemplo, para el adjetivo claro: "turbio" (agua clara), "oscuro" (color oscuro), "complicado" (un asunto claro) (Lehmann y Martin-Berthet, 1998).

antónimos se componen siempre de semas[25] comunes: así hermano y hermana comparten los semas /ser humano/ /nacido de los mismos padres/, y se oponen por el sema /relativo al sexo/. En consecuencia, la relación de antonimia se puede definir como la relación que une dos palabras de la misma categoría gramatical que contienen una parte de sus sememas en común y otra que se opone. La antonimia, además, se compone de diferentes tipos de oposiciones, principalmente binarias. Se distinguen generalmente tres tipos de antónimos: (1) *antónimos contradictorios o complementarios*, expresan una relación de disyunción exclusiva, es decir, la negación de una de las palabras acarrea la afirmación de la otra y los dos términos no se pueden negar simultáneamente, por ejemplo, hombre/mujer, presente/pasado. (2) *Antónimos contrarios o gradables*, definen los extremos de una escala de gradación implícita y autorizan la existencia de grados intermedios. Grande/ pequeño, calor/ frío, amor/ odio[26]. (3) *Antónimos recíprocos*, son aquellos que obligan a una sustitución de uno por el otro, en un enunciado dado, para conservar la relación, por ejemplo, médico/enfermo, padre/hijo.

La *relación de inclusión* principal es la generalización o *hiperonimia* y su inversa la especialización o *hiponimia*[27]. Un ejemplo de relaciones hiperonimia-hiponimia se muestra en la figura 2.2. Las relaciones de hiperonimia-hiponimia son asimétricas y transitivas por lo que "ordenan" las palabras en jerarquías, simples o múltiples, de significados más generales a más específicos. En general las estructuras son reticulares[28], pero se utiliza el término jerarquía para enfatizar la dependencia conceptual de los hipónimos de los hiperónimos.

[25] Sema. Unidad mínima de significado lexical o gramatical (DRAE, 2001).

[26] Dos propiedades les distinguen de los contradictorios: están sujetos a la gradación y la negación de uno de ellos no implica obligatoriamente la afirmación del otro.

[27] Considerando que pueden distinguirse varios tipos de relaciones de generalización/especialización (hiperonimia/hiponimia, es-un, es-un-tipo-de, instancia, subsunción (Cruse, 2002).

[28] Forman estructuras matemáticas denominadas retículos. Un retículo es un conjunto, de términos en este caso, y una relación de orden que los organiza. Cumple una serie de propiedades y se visualizan gráficamente utilizando los diagramas de Hasse (Hortalá et al., 2001).

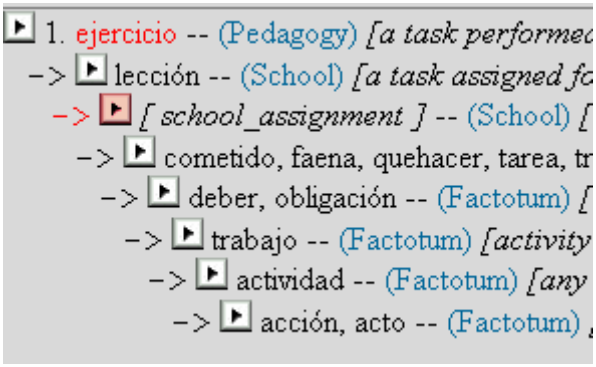


Figura 2.2. Hiperónimos e hipónimos del primer significado de la palabra "ejercicio"[29]

La *co-hiponimia* se establece, dentro de una misma jerarquía, entre las palabras relacionadas con un mismo hiperónimo, por ejemplo, "tulipán" y "margarita" son co-hipónimos de "flor"; "primavera", "verano" son co-hipónimos de "estación". Los co-hipónimos son unidades del mismo rango porque están situadas al mismo nivel de la relación que las une a la hiperonimia. Se diferencian entre ellas por una o más características específicas. Contrariamente a lo que pasa en la relación de antonimia fundada en una oposición binaria, la negación de uno de los co-hipónimos no implica necesariamente la afirmación del otro co-hipónimo, la elección permanece abierta: si x no es un "tulipán", puede ser una "margarita", una "rosa". Los co-hipónimos son mutuamente excluyentes: una "flor" es o un "tulipán" o una "rosa" o una "margarita". Además, los co-hipónimos pueden mantener entre ellos relaciones de sinonimia o de antonimia. "Vivaracho" y "jovial" se pueden considerar co-hipónimos de "alegre" y pueden pasar por sinónimos; contrariamente "comprar" y "robar" co-hipónimos de "conseguir" son antónimos. Además, un mismo par léxico puede cambiar de estatus léxico según el contexto: "soltero" y "casado" son antónimos complementarios que se convierten en co-hipónimos en el marco de un formulario de estado civil, junto a "divorciado", "viudo" (Lehmann y Martin-Berthet, 1998).

La relación *parte/todo u holonimia/meronimia* también es una relación jerárquica (figura 2.3.) orientada y transitiva en la que uno de los términos denota una parte y el otro el todo (relativo a esa parte)[30]. La diferencia con la relación de hiponimia-hiperonimia es que es una relación de pertenencia en vez de inclusión por lo que los merónimos no heredan los atributos de los homónimos. Las relaciones de dependencia meronímica son variadas y complejas: miembro/ conjunto (árbol/ bosque), componente/

[29] Fuente: multiwordnet en http://multiwordnet.itc.it/online/multiwordnet.php

[30] Únicamente los nombres que refieren a referentes divisibles y discretos son susceptibles de ser merónimos.

ensamblaje (asa/ taza), porción/ masa (parte/ tarta), material/ objeto (acero/ bicicleta) (Lehmann y Martin-Berthet, 1998). Por ejemplo en el vocabulario WordNet[31] se utilizan tres tipos de relaciones holonímicas: miembro, sustancia y parte (Miller, 1995).

Además de las relaciones léxicas "clásicas", los vocabularios pueden incluir muchas otras relaciones asociativas como, por ejemplo, familia (fuego/bombero), agente (estudiante-aprender), instrumento (practicar-ejercicio) o localización (estudiante-facultad)[32].

Como veremos en el siguiente capítulo y siguientes secciones, las relaciones semánticas, especialmente las de sinonimia e hiponimia-hiperonimia, son básicas para construir los tesauros. Los tesauros son un tipo de vocabulario que se caracteriza por incluir estas relaciones, bien implícitamente –Tesauro de Roget- (ver figura 2.20 en sección 2.6.3), o bien explícitamente -tesauros utilizados para clasificación, indexación de información, documentos o recursos didácticos digitalizados (ANSI/NISO Z39.19, 2005)- (ver figuras 2.23, 2.24 y 2.25 en la sección 2.6.3).

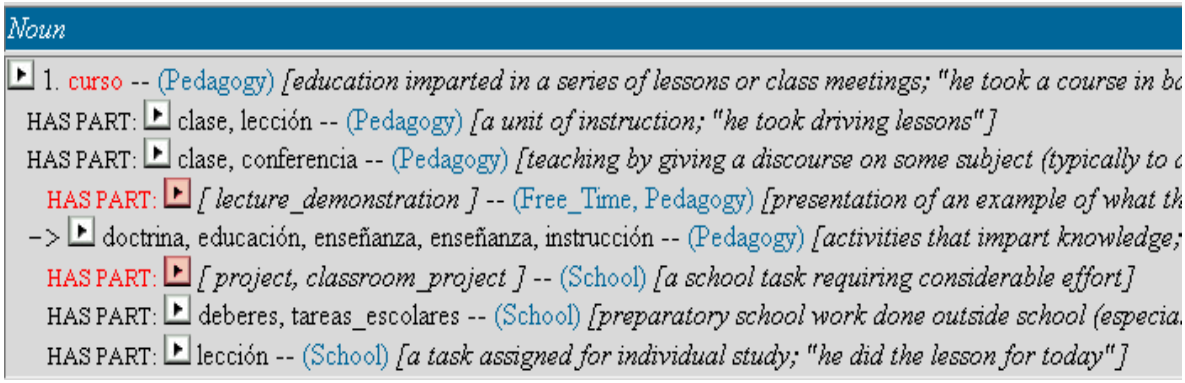


Figura 2.3. Merónimos del término curso, respecto de su primer significado en la red semántica WordNet.

2.4. Los vocabularios en los sistemas de recuperación de la información

En el contexto de la RI los vocabularios son sistemas de organización del conocimiento cuya finalidad es incrementar la efectividad de los procesos de recuperación y exploración de los datos, información u objetos gestionados por dichos sistemas. Constituyen un *"puente"* entre la concepción abstracta del dominio que tienen los

[31] http://wordnet.princeton.edu/

[32] Estas relaciones constituyen el fundamento para el procesamiento semántico de frases en lenguaje natural basado en formalismos como la gramática de casos (Fillmore, 1968), la gramática de dependencia conceptual (Schank y Tesler, 1969) o las redes semánticas (Sowa, 1991).

humanos y la organización real física de los datos en los ordenadores y entre las representaciones en lenguaje natural y las representaciones codificadas de los lenguajes de programación (Lancaster, 1986; CEN CWA 14871, 2003; Aitchison y Clarke, 2004).

La efectividad de un sistema RI se mide, normalmente, mediante los indicadores de exhaustividad[33] y de precisión. La *exhaustividad* se refiere a la capacidad de recuperar u obtener de un repositorio o base de datos los objetos (documentos o recursos) relevantes para el usuario en un proceso de consulta o búsqueda[34]. La precisión se refiere a la capacidad de obtener el mínimo número de objetos no relevantes en un proceso de búsqueda[35]. La exhaustividad y la precisión son parámetros inversamente proporcionales: mayor exhaustividad implica menor precisión y viceversa.

Aunque cada uno de los principales componentes de un sistema RI -el sistema de indexación, el vocabulario el sistema de búsqueda y la interfaz del usuario- tiene una gran influencia en la efectividad del sistema, en este trabajo nos centramos únicamente en los factores que dependen sólo del vocabulario y que, por lo tanto, deben de tenerse en cuenta para su construcción y actualización. Las cuestiones que hay que tratar son dos: cómo puede el vocabulario mejorar la efectividad de los procesos de indexación y búsqueda en la RI y viceversa, cuáles son las fuentes de fallos atribuibles a los vocabularios que merman la efectividad de la recuperación.

Las dos fuentes principales de fallos atribuibles a los vocabularios provienen de la especificidad del vocabulario y de las relaciones ambiguas o espurias (Lancaster, 1986). Además, la especificidad es probablemente el factor que más influye en obtener buenos resultados para la exhaustividad o la precisión. Cuando un tesauro es muy específico permite describir un objeto con muchos términos o categorías pequeñas; esto significa una mayor precisión en la indexación pero, al mismo tiempo, complica la localización de los objetos, porque el usuario tiene que tener un conocimiento del dominio profundo para poder expresar con suficiente precisión la consulta. Por el contrario, si el tesauro es general, el usuario tiene más probabilidades de encontrar los documentos o recursos que busca, utilizando conceptos de significado amplio, aunque los resultados obtenidos

[33] En la bibliografía se utiliza de forma extendida el término en inglés, recall.

[34] Cuantitativamente es la razón entre el número de objetos relevantes recuperados y el número total de objetos relevantes almacenados en el repositorio (y multiplicado por 100 si se expresa en tanto por ciento). Por ejemplo una exhaustividad (recall) de 1/3 (ó 33%) significa que se recupera uno de cada 3 documentos relevantes para el usuario.

[35] Se mide numéricamente como el cociente entre el número de objetos relevantes obtenidos y el número total de objetos obtenidos en un proceso de consulta o búsqueda (y multiplicado por 100 si se expresa en tanto por ciento). Una precisión de 7/50 (14%) significa que de 50 objetos encontrados sólo 7 son relevantes para el usuario.

serán, en su mayoría, irrelevantes (figura 2.4). En definitiva, los términos deben ser lo suficientemente específicos como para permitir la recuperación de objetos de contenido deseados, minimizando los resultados no deseados pero, al mismo tiempo, debe contener términos suficientemente generales como para recoger consultas de usuarios no expertos en el dominio –por ejemplo los alumnos durante su aprendizaje.

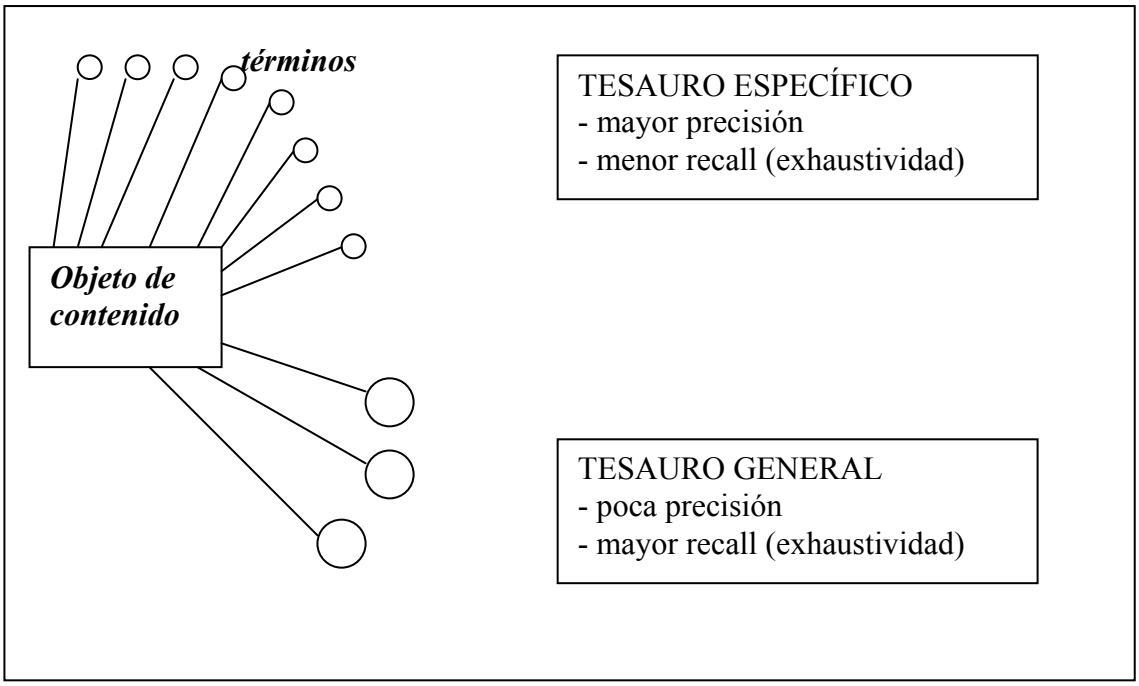


Figura 2.4. Relación entre la especificidad de un tesauro y la precisión y exhaustividad de la RI (Lancaster, 1986)

La segunda fuente de fallos procede de las relaciones entre los términos en los casos de sinonimia no controlada y relaciones ambiguas entre términos. El control de la sinonimia es necesario para hacer coincidir el lenguaje de indexación del experto con lenguaje de consulta del usuario. La ambigüedad de las relaciones aparece en los sistemas postcoordinados que indexan los objetos utilizando múltiples términos sin coordinar, es decir, sin sintaxis. Se producen fallos del tipo *coordinaciones falsas*[36] de términos o *relaciones incorrectas*[37] que sólo pueden corregirse con una mayor pre-

[36] Por ejemplo, un documento que trata sobre "motores para una fusión nuclear limpia" se indexa con los términos *motor, fusión, nuclear*, *limpieza*. Si en la consulta se busca documentación sobre la "limpieza de motores" la coordinación falsa entre "motor" y "limpieza" en el documento indexado (la coordinación correcta sería entre "motor" y "fusión") hará que éste se obtenga como resultado erróneo de la consulta.

[37] En este caso los términos están relacionados tanto en la indexación como en la consulta, pero no de la misma forma. Por ejemplo, supongamos de nuevo un documento sobre "Diseño de aviones por ordenador" y una consulta que se refiera a "diseño de ordenadores". En ambos casos los términos "diseño" y "ordenador" está relacionados, pero con distintas funciones argumentales. En el primer caso, "ordenador" es el argumento "instrumento" de *diseño*, mientras que en el segundo tiene la función de "objeto".

coordinación o con cualificadores para la desambiguación[38], lo cual incrementa la especificidad del vocabulario y los costes de la indexación.

Los sistemas de RI que incorporan vocabularios mejoran la exhaustividad de la búsqueda porque controlan las relaciones entre un término de la consulta y cualquiera de los términos relacionados con él en el vocabulario. De esta forma, se pueden encontrar todos los posibles términos relacionados por la forma (control de forma) o el significado (control de sinónimos, cuasi-sinónimos, control de hipónimos-hiperónimos, holónimos-merónimos y otros términos asociados). Los dispositivos para incrementar la precisión son entre otros, la coordinación[39], homógrafos y notas de ámbito[40] y la frecuencia[41]. Los vocabularios se utilizan en los procesos de indexación de la información o recursos que almacenan y para búsqueda y exploración[42]

2.4.1. El vocabulario en la indexación

La indexación con vocabulario es el proceso de representación de la información o del objeto digital (su contenido, contexto, estructura, derechos de propiedad y otros) con términos procedentes del vocabulario controlado -indexación por asignación- o con términos extraídos de los propios objetos -indexación por extracción- (figura 2.5).

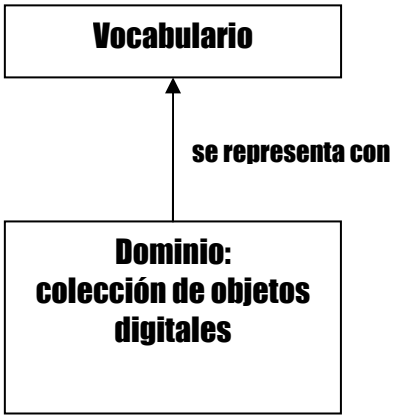


Figura 2.5. El vocabulario como sistema de representación de un dominio

[38] Por ejemplo, añadiendo a los términos un número o etiqueta idéntica en los términos coordinados tanto en la indexación como en la consulta o bien añadiendo, para el segundo caso, la función de los términos (ver capítulo 15 de Lancaster, 1986).

[39] La coordinación es el mecanismo más importante para mejorar la precisión. Consiste en combinar términos bien durante el indexado de los objetos o recursos digitalizados o bien durante la búsqueda para afinar una consulta.

[40] Son mecanismos para desambiguar, restringir y clarificar los significados.

[41] Permite diferenciar estadísticamente los términos más probables basándose en el uso, el contexto o incluso el perfil del usuario. En algunos sistemas se visualiza mediante el diferente tamaño de letra de los términos.

[42] En Aitchison et al., 2000 se presenta una tipología de tesauros en base al criterio de uso en los Sistemas de RI: tesauros de indexación-búsqueda, tesauros únicamente de indexación, tesauros únicamente de búsqueda y tesauros con otros usos.

En el caso de la indexación por extracción, el vocabulario y los índices se crean durante el proceso de indexación, mientras que en el caso de asignación, sólo se generan índices que reproducen la categorización de los objetos respecto de un vocabulario ya existente (Lancaster, 1986). En cualquiera de los casos, y desde la perspectiva de vocabularios para la RI, la indexación consiste en escoger un conjunto pequeño y limitado de términos *que representen el significado del objeto* o el significado del contenido del objeto y asociarlos a dicho objeto. Esta representación semántica abre otra vía para la recuperación -en función de su contenido, de su significado- utilizada por los algoritmos de búsqueda para mejorar la exhaustividad y la precisión y por las interfaces de acceso a los objetos almacenados para mostrarle al usuario, con los términos organizados del vocabulario, una mapa de contenidos para navegar.

La indexación se realiza en tres fases: (1) el análisis del dominio (colección de objetos), que conlleva la extracción de los conceptos claves de cada uno de los objetos; (2) la selección de los términos de un vocabulario controlado más representativos para describir a los objetos; y (3) la creación de estructuras de datos (índices) para almacenar las asociaciones entre objetos y términos (figura 2.6).

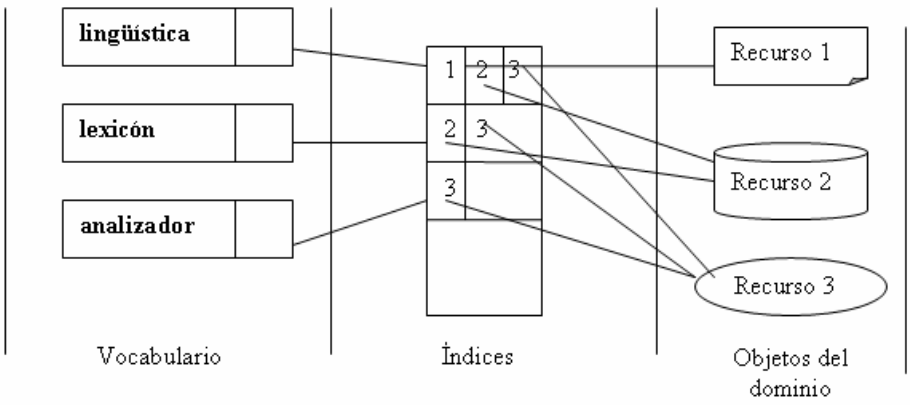


Figura 2.6. Indexación con vocabularios

Este proceso de indexación puede ser manual -en cuyo caso también se denomina *indexación inteligente*-, automático o semiautomático (Aitchison et al., 2000). En el primer caso, un experto (o comité de expertos) analiza el contenido, el contexto, y/o la estructura de un recurso y le asigna un conjunto de términos a partir de un vocabulario

controlado[43] (categorización por asignación) o a partir de la información y metainformación del propio recurso (categorización por extracción). El uso manual de un vocabulario para indexar permite obtener un alto nivel de exactitud en la descripción de los recursos, la capacidad de incorporar el significado contextual en la descripción y la posibilidad de indexar objetos no textuales (imágenes, aplicaciones, etc.). Sin embargo, tiene como inconvenientes el coste elevado y, en muchos casos, la falta de consistencia y exhaustividad en las descripciones (Bennett, 2002)[44].

La indexación automática se fundamenta en algoritmos que analizan estadísticamente las secuencias de palabras de los objetos de contenido (de los metadatos o de los datos textuales: títulos, resúmenes o texto completo del objeto)[45]. En la indexación por extracción automática se identifican patrones de comportamiento de las palabras a partir de variables como la frecuencia de uso, colocación, orden, proximidad, etc. Con ello se identifican las palabras y relaciones que mejor representan el contenido del objeto[46]. En la indexación automática por asignación las palabras extraídas se comparan con los términos de un vocabulario controlado para seleccionar aquellos con mayor similitud como descriptores del objeto. El resultado son agrupaciones ("clusters") de objetos de contenido que muestran patrones de comportamiento similares, etiquetados mediante la secuencia de términos extraídos de los vocabularios controlados o de los propios objetos, y que son los que mejor representan su contenido (Lancaster, 1986; Centelles, 2005).

Los ejemplos de sistemas de indexación automática que utilizan o generan vocabularios de indexación son numerosos, como muestra, citamos los trabajos de Montejo, 2001; Gómez Hidalgo et al., 2004; Mao y Chu, 2007 o los sistemas comerciales SPIRIT[47] e IDOL Server[48] (figura 2.7).

---

[43] En este caso entendemos que el vocabulario puede ser también un lenguaje (vocabulario + reglas sintácticas) y se trata de vocabularios para sistemas precoordinados.

[44] Bennett presenta algunos ejemplos sobre los altos costes: Yahoo utiliza 200 personas para la indexación de páginas web según su taxonomía de 500.000 términos; MEDLINE (la biblioteca nacional de medicina) gasta 2 millones de dólares al año para la indexación de los artículos con el tesauro MeSH.

[45] En el momento actual, los algoritmos diseñados para el análisis de frecuencias, utilizan alguno o una combinación de varios de los siguientes métodos de análisis de secuencias de palabras: métodos probabilísticos (método bayesiano, método de Rocchio...); métodos vectoriales (método K-Nearest Neighbor, Support Vector Machines...); y árboles y listas de decisión (Centelles, 2005).

[46] Se pueden formar automáticamente distintos tipos de agrupaciones de términos relacionados a partir de la co-aparición de los mismos en los documentos analizados (Salton y McGill, 1986).

[47] SPIRIT es un indexador automático con un motor de búsquedas inteligente y que utiliza el lenguaje natural. http://www.spiritengine.com/. Permite la indexación automática de una gran diversidad de tipos de información (archivos .doc, .pdf, .rtf. html, contenido de sitios Intranet o Internet, información proveniente de bases de datos, etc.) y contiene un módulo de indexación automática que permite utilizar una taxonomía estándar o definir una personalizada. La búsqueda se hace mediante el navegador y la

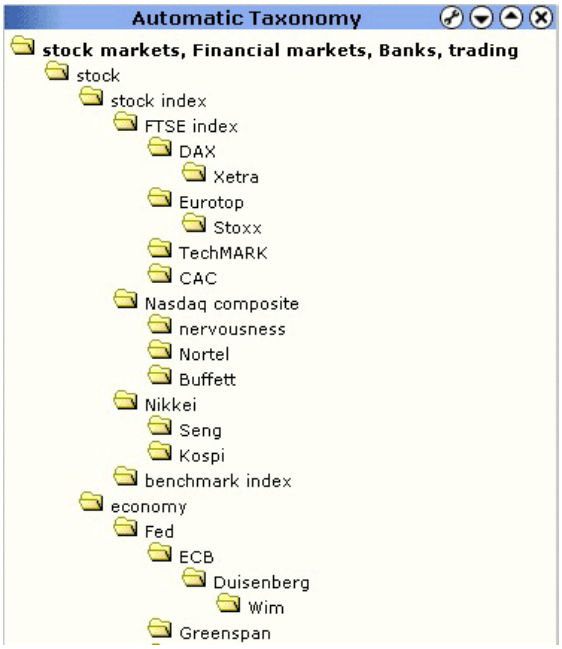


Figura 2.7. Generación automática de taxonomías para la indexación de fuentes de información o de colecciones de recursos[49]

La indexación automática es más rápida, menos costosa y más objetiva que la manual, pero el nivel de exactitud de las descripciones es más bajo y suele necesitar cierta intervención humana para corregir los resultados obtenidos.

Los sistemas de indexación semiautomática o híbrida combinan la inteligencia humana, que puede identificar los diferentes significados de los objetos, y la eficiencia de los automatismos. Se pueden identificar cuatro familias de sistemas semiautomáticos de indexación (Centelles, 2005).

- Sistemas que analizan estadísticamente los recursos y presentan a los expertos humanos términos recomendados de indexación para que éstos los revisen y aprueben. Un ejemplo de este tipo de sistemas es Ultraseek Advanced Classifier (http://www.verity.com/products/ultraseek/index.html).
- Sistemas basados en reglas de búsqueda. Permite la opción avanzada de vincular a los términos de un vocabulario ecuaciones de búsqueda diseñadas por especialistas. Mediante un algoritmo, el sistema analiza los metadatos de los objetos o los documentos de texto y determina cuál o cuáles son las

tecnología de análisis lingüístico permite a los usuarios formular las preguntas en lenguaje natural utilizando frases usuales. Se obtienen así los documentos buscados mediante un conjunto de términos utilizados por los creadores de documentos, teniendo en cuenta los sinónimos y las expresiones con un sentido similar.

[48] IDOL Server http://www.autonomy.com/content/Products/IDOL incluye un módulo de categorización automática y de generación automática de taxonomías que se basa en el método probabilístico bayesiano.

[49] Fuente: http://www.autonomy.com/Media/Content/IDOL/Slots/Taxonomy/01

ecuaciones con las que manifiesta mayor coincidencia. A continuación, asigna el documento a los términos del vocabulario que tienen vinculadas dichas reglas de búsqueda. Son ejemplos de este tipo de sistemas K2 Enterprise (http://www.verity.com/products/k2_enterprise/index.html) y Ultraseek Content Classification Engine (http://www.verity.com/products/ultraseek/cce.html), ambos de Verity.

- Sistemas de indexación basados en agrupamiento ("clustering") y en colecciones de documentos de entrenamiento o ejemplares. Inicialmente se indexa de forma manual una colección limitada de documentos que son considerados como los más relevantes por los especialistas. El sistema utiliza esta indexación inicial como base para analizar y agrupar nuevos documentos, aplicando algoritmos de similitud entre los nuevos documentos y los documentos ya indexados. Los nuevos documentos se indexan con el conjunto de términos procedentes de los documentos más cercanos a ellos. Un ejemplo de este tipo de sistemas es Mohomine Classifier (http://www.kofax.com/products/mohomine/classifier.asp), de Mohomine.
- Sistemas de indexación basados en el análisis lingüístico. Un ejemplo de este tipo de sistemas es Inxight Smart Discovery http://www.bitpipe.com/plist/term/Indexing-(Information-Management).html

En los sistemas de indexación semiautomáticos el proceso está guiado por el razonamiento humano, pero tienen la capacidad de acumular experiencia y aprender, lo que los hace cada vez más eficientes. Entre los puntos débiles, cabe destacar la exigencia de conocimientos, habilidades y esfuerzos de gestión y mantenimiento.

En la indexación, el uso de vocabularios "recomendados", estándares y conocidos aporta a los sistemas de RI una ventaja nada despreciable: la interoperabilidad que permite reutilizar y compartir la información y objetos que gestionan. Esta es la razón por la que se recomienda fehacientemente el uso de vocabularios ya existentes, el uso de estándares para la construcción de nuevos vocabularios y el registro de vocabularios (CEN CWA 14871, 2003; ANSI/NISO Z39.19, 2005; Nilsson et al., 2009; IMS Digital Repositories, 2003).

Otro aspecto que hay que destacar sobre el uso de vocabularios en la indexación para la RI es su combinación con los metadatos. Los metadatos, como se verá en el capítulo siguiente, constituyen uno de los métodos de descripción más eficaz que existe en la actualidad. A pesar de que no existe un modelo de metados único, el método para

describir los objetos utilizando vocabularios es el mismo. Consiste en (1) utilizar los términos de uno o más vocabularios para dar valores a las propiedades de los metadatos de los documentos; y (2) declarar explícitamente los vocabularios que se han utilizado. El esquema de metadatos de carácter general más utilizado es el Dublin Core (Powel et al., 2005), y en el dominio del *e-learning* es LOM (figura 2.8).

```
<html>
<head><title>Introducción a la programación en Prolog</title>
<meta name="DC.Type" content="text">
<meta
       name="DC.Subject"
       scheme="UCD"
       content="Language.Linguistics.Literature/Linguistics and
       Language/Languages/Artificial Languages">
<meta name="DC.Description" content="Este documento es una
       introducción al lenguaje de programación lógica Prolog.
       Constituye material docente de la asignatura Lingüística
       Informática de la titulación de Lingüística de la Facultad
       de Filología de la UCM">
<meta name="DC.Creator" content="Ana Fernández-Pampillón">
<meta name="DC.Title" content=" Introducción a la programación en
       Prolog">
<meta name="DC.Identifier" scheme="URI"
       content="https://campusvirtual.ucm.es/linguistica-
       informatica/intro.html">
…….
```

Figura 2.8. Uso del vocabulario UDC (sistema de clasificación universal decimal) en los metadatos Dublín Core. El atributo tema (DC.Subject) toma como valores un conjunto de términos relacionados (hiperonimia) del vocabulario[50].

La utilización de vocabularios en la descripción e indexación de objetos de contenido tiene las ventajas de los lenguajes controlados (Aitchison et al., 2000): (i) el tratamiento de los aspectos semánticos y sintácticos del lenguaje; (ii) la representación de conceptos implícitos; (iii) la creación de una visión global de los dominios que son objeto de representación; (iv) la exhaustividad en la indexación; y (v) la solución de los problemas que conllevan los contextos multilingües. Desde el punto de vista de los sistemas de RI el uso de vocabularios ofrece dos importantes beneficios adicionales:

- Por un lado, rentabiliza los esfuerzos de construcción y mantenimiento del vocabulario y de la indexación de recursos porque esta misma herramienta –

[50] El uso de vocabularios en metadatos no se limita a los atributos que expresan el contenido de los objetos, del tipo *materia*, *tema* o *disciplina*. Otros atributos como los relativos al contexto y a la estructura de los objetos también pueden ser expresados mediante términos extraídos de un vocabulario.

el vocabulario- puede ser reutilizada en el desarrollo de sistemas de búsqueda, navegación, personalización, etc.

- Por otro lado, permite mantener la consistencia conceptual y designativa en la representación de los elementos de un mismo dominio, lo cual crea en los usuarios una imagen de consistencia en la colección de recursos.

2.4.2. El vocabulario en la búsqueda y navegación

Los sistemas que permiten buscar contenidos y recursos en un entorno web pueden clasificarse en tres grandes tipos: de exploración ("browsing"), de recuperación ("searching") y de filtrado ("filtering") (Centelles, 2005).

1. Los *sistemas de búsqueda por exploración* ofrecen a los usuarios una estructura organizada de términos, el vocabulario, donde se incorporan los recursos de información y un mecanismo de navegación por dicha estructura para localizar los recursos relevantes para sus necesidades de información. La forma en que se presenta el vocabulario varía desde la simple navegación en una lista alfabética y búsqueda restringida a los términos del vocabulario (figura 2.9) hasta los sistemas híbridos que permiten ir seleccionando términos durante la navegación en un vocabulario –con todos los términos y relaciones-, y combinar estos términos en una expresión de búsqueda (Dalmau et al., 2005).

Figura 2.9. Presentación del tesauro TGM (Thesaurus for Graphical Material[51]) para la exploración de las imágenes de la colección fotográfica "US Steel Gary Works Photograph Collection, 1906-1971[52]"

[51] http://www.loc.gov/rr/print/tgm1/toc.html

2. Los *sistemas de recuperación de información* ofrecen a los usuarios la posibilidad de crear una expresión de búsqueda a partir de una palabra o una combinación de palabras (figura 2.10). Estos sistemas de exploración son especialmente convenientes para situaciones de búsqueda en las que los usuarios pueden concretar con suficiente detalle la necesidad de información (búsqueda de un objeto conocido). El vocabulario se incorpora al sistema de recuperación para auxiliar al usuario en la identificación de términos relevantes para la creación de la expresión de búsqueda, y también para mejorar los procesos de presentación de resultados y reformulación de la consulta. Los sistemas de exploración y de recuperación suponen la interacción a tiempo real entre el usuario y el mecanismo de búsqueda.

Figura 2.10. Vocabulario (tesauro ETB) para la selección de términos de búsqueda en el repositorio AGREGA[53]

3. La tercera modalidad, *los sistemas de filtrado*, permite definir unas preferencias de información en el perfil del usuario y recibir una respuesta automática cada vez que el sistema identifica recursos relevantes para dicha preferencia. En este caso, el tesauro permite al usuario seleccionar términos relevantes para la concreción de su perfil.

[52] http://www.dlib.indiana.edu/collections/steel/index.html
[53] Buscador de recursos educativos del repositorio AGREGA: http://www.proyectoagrega.es/default/home.php

En cualquiera de estos tipos, el vocabulario tiene, en el proceso de búsqueda, dos funciones básicas: (i) define un lenguaje completo y preciso para hacer consultas, y (ii) guía al usuario, o al sistema, en la búsqueda ofreciéndole todas las posibles alternativas, con los términos y las relaciones entre términos, para recuperar el contenido u objeto deseado. Además, en el proceso de búsqueda se llevan a cabo dos operaciones: (i) el análisis de la consulta, y (ii) su traducción al lenguaje de representación de los objetos almacenados. Si el usuario utiliza para expresar la consulta el mismo lenguaje que el utilizado para la indexación, el proceso de búsqueda se simplifica porque el análisis de la consulta supone únicamente la comprobación de que la secuencia de términos pertenece al vocabulario[54] y la traducción se reduce a un ajuste de patrones entre la secuencia de entrada y las secuencias de indexación de los objetos almacenadas en los índices. El inconveniente, sin embargo, es que al usar un vocabulario limitado el usuario no siempre es capaz de expresar lo que realmente necesita (Lancaster, 1986). Para resolver esta cuestión se presentan dos aproximaciones: (i) expresar la consulta en lenguaje natural, lo cual es sencillo para el usuario, y aplicar técnicas de procesamiento del lenguaje natural para analizar y traducir esta consulta; y (ii) actualizar el vocabulario, automática o manualmente, para que se vaya acercando al vocabulario del usuario. Esto supone aplicar una metodología de actualización que incluya los criterios para aceptar nuevos términos y relaciones sin perder la consistencia del vocabulario (Aitchinson et al., 2000).

El problema que se plantea es que el vocabulario utilizado por los indexadores y el vocabulario utilizado por los usuarios para expresar una misma idea no es el mismo. Todavía es más complicado si, además, participa un tercer grupo: los autores de los objetos de contenido que los crean y los describen utilizando, también, sus propios términos. *Por eso es necesario replantearse, en ciertas situaciones, este enfoque, promovido por los estándares, de construir y utilizar vocabularios generales "autorizados". Cuando el papel de autor, indexador y usuario lo realiza el mismo grupo de personas, no parece recomendable utilizar vocabularios "externos" no ajustados al lenguaje de los autores y usuarios, sino utilizar un vocabulario común y compartido por el grupo de autores y usuarios.* La necesidad de crear y utilizar

[54] En el caso de sistemas pre-coordinados es sencillo porque se dispone de unas reglas de combinación; sin embargo, en los sistemas post-coordinados hay que calcular si los términos están relacionados y cómo. Esto puede dar lugar a los fallos de coordinación falsa y relaciones incorrectas comentados en la sección anterior.

vocabularios ajustados al dominio de conocimiento que representan y al lenguaje de los usuarios en el contexto académico es la motivación de este trabajo de tesis.

2.5 Los vocabularios en la explotación didáctica de recursos digitalizados

El uso de los vocabularios en la descripción y recuperación de información data de los años 50. Con la RI se comienza a desarrollar y aplicar *índices de coordinación,* que son listas de términos interrelacionados semánticamente y/o sintácticamente, que se pueden combinar, coordinar, durante la indexación -precoordinación- o durante la búsqueda –postcoordinación. A finales de los 60 existen ya sistemas de RI que utilizan estos índices como herramientas de indexación y búsqueda de objetos de información (Austin, 1976).

La RI se basa en utilizar técnicas que procesen descripciones de la información, es decir, que procesen la metainformación. De la misma forma, se aborda la recuperación de recursos digitalizados en la Web o en repositorios de gran tamaño: utilizando esas descripciones, primero se describe el recurso y luego se busca. En la actualidad se puede considerar que existen tres aproximaciones para describir y recuperar recursos digitalizados. La primera consiste en asociar metadatos a cada recurso. Los metadatos son un conjunto de propiedades y valores que describen un recurso desde múltiples puntos de vista –contenido, autoría, formato, propiedad intelectual, propósito, etc.- (figura 2.11). La búsqueda de dichos objetos, *búsqueda basada en propiedades,* se lleva a cabo a partir de sus propiedades y valores, y es muy eficiente. El inconveniente que tiene es la falta de homogeneidad en las propiedades y en los valores de las propiedades. Esto afecta a la efectividad la búsqueda y a la reutilización e interoperabilidad de recursos entre los distintos repositorios.

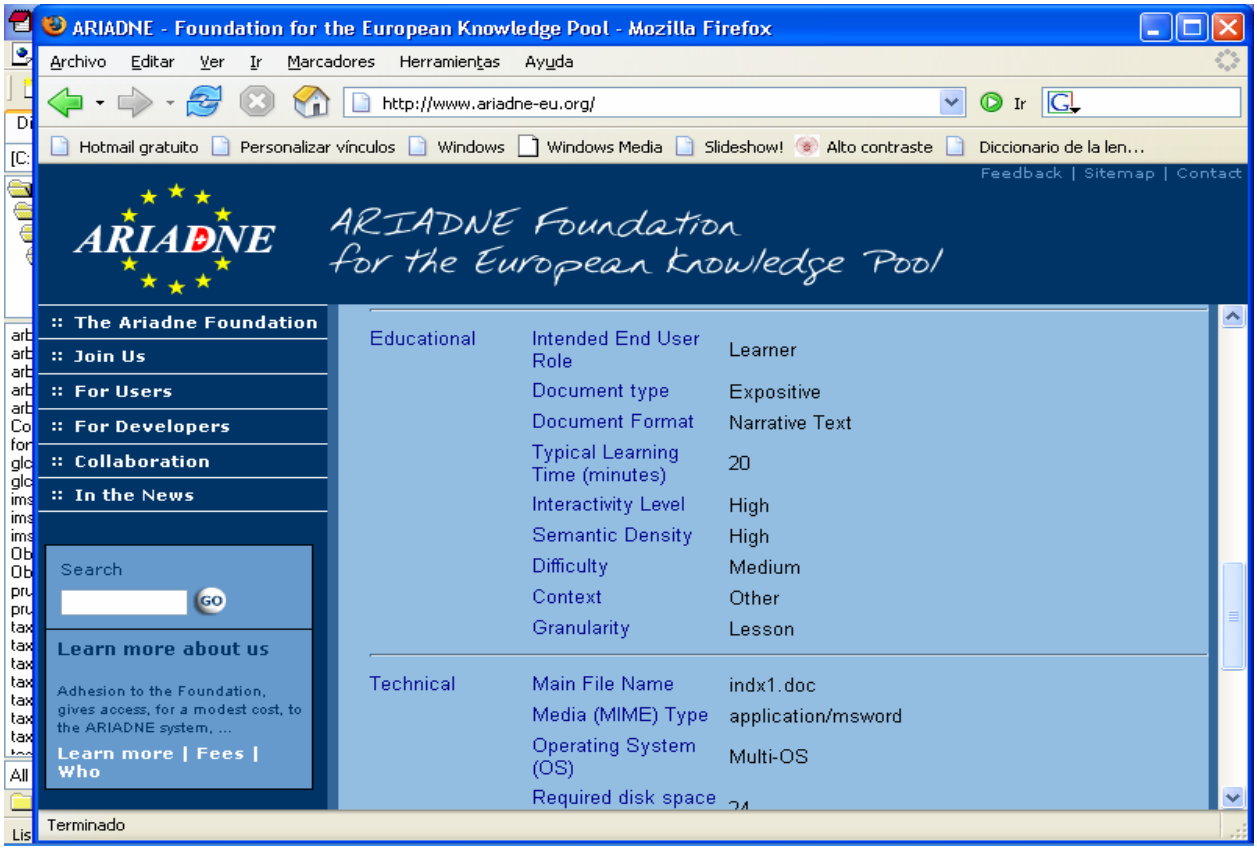


Figura 2.11. Metadatos de un recurso educativo en el repositorio Europeo ARIADNE

La segunda alternativa es la descripción y búsqueda de recursos utilizando sistemas de indexación, también llamados lenguajes de indexación (Slavic, 2000). Un *sistema de indexación* (*lenguaje de indexación*) es un vocabulario, y en algunos casos un conjunto de normas sintácticas[55]. Además, se distinguen dos tipos de lenguajes de indexación: a) los que usan términos o palabras del lenguaje natural (vocabularios controlados), y b) los que usan símbolos (clasificaciones bibliográficas y sistemas de representación del conocimiento). En RI se utilizan vocabularios controlados que contienen términos en lenguaje natural, o categorías conceptuales en un pseudos-lenguaje natural, por ejemplo, las ontologías, que son fácilmente interpretables por las personas (Woods et al., 2000). Estos términos, si son indexados por los sistemas de almacenamiento, gestión y recuperación de información, como los repositorios, bases de datos, bases de conocimiento, permiten acceder a los objetos buscados por su contenido expresado en el propio lenguaje del usuario -o en un lenguaje cercano, lo que mejora la efectividad de los procesos de búsqueda (Aitchison et al., 2000).

[55] Los lenguajes de indexación precoordinados describen los objetos mediante combinaciones de términos fijadas en el proceso de indexación. Estos lenguajes, por lo tanto, además del vocabulario necesitan de un conjunto de normas sintácticas que normalicen la combinación de términos en la indexación.

Al mismo tiempo, los vocabularios proporcionan un modelo conceptual del dominio de conocimiento que permite presentar al usuario la organización conceptual de los objetos del dominio para poder entender sus contenidos y localizar, mediante la navegación, de forma más precisa, los objetos con el contenido deseado. Esta es una de las aplicaciones más efectivas que se está incorporando en los sistemas de almacenamiento y recuperación de información, en las bases de datos, los sistemas de navegación Web y en otros entornos que precisen de funciones de búsqueda e identificación de contenido mediante algún tipo de descripción en lenguaje natural (ANSI/NISO Z39.19, 2005).

Sin embargo, el problema de utilizar únicamente vocabularios para indexar y buscar objetos es que sólo permite describir un aspecto de la colección de objetos, normalmente su contenido temático, y no se documentan otras características o propiedades que podrían utilizarse para precisar la búsqueda. Además, los vocabularios presentan los problemas de ambigüedad y polisemia del lenguaje natural que también afectan a la eficacia del proceso de búsqueda.

La tercera alternativa es el uso conjunto de metadatos y vocabularios. Las últimas versiones de los estándares de metadatos proponen autodocumentar los objetos de contenido utilizando esquemas de metadatos en los que ciertas propiedades toman como valores posibles los términos de uno o varios vocabularios referenciados y recomendados en el mismo estándar (Dublin Core, 2008; IEEE-LOM 1484.12.1, 2002). Este conjunto de propiedades y vocabularios que describen los recursos de una forma común y compartida, al menos, en cada comunidad de usuarios, resuelve el problema de la compatibilidad y permite el intercambio, uso compartido e integración de colecciones de objetos[56], por ejemplo, ARIADNE, GEM. Además, mejora la eficacia de la búsqueda porque permite combinar una búsqueda basada en propiedades con una búsqueda basada en términos del lenguaje natural. Finalmente, permite incorporar una búsqueda basada en la exploración del vocabulario que representa el dominio de conocimiento de las colecciones de objetos de contenido para ayudar al usuario a buscar y seleccionar los objetos que más se acercan a la idea que se consulta.

[56] En este sentido, merece la pena destacar el papel que juegan actualmente los vocabularios como mediadores en los procesos de compartir e intercambiar recursos entre personas y sistemas diferentes, por ejemplo, entre plataformas e-learning. Un mediador es un módulo de software que emplea el conocimiento codificado para crear información en una capa superior a las aplicaciones que integra y conecta (Aguirre et al., 2004).

Los vocabularios previstos por los estándares de metadatos deben cumplir con una serie de requisitos: autoridad[57], estabilidad, mantenimiento, difusión, cobertura, multilinguísmo, aplicabilidad a las necesidades de los usuarios, grado de conformidad con los estándares y especificaciones (IMS Meta-data, 2004; CEN CWA 14871, 2003). No se permite el uso de folksonomías[58], ni vocabularios particulares, ad hoc. Se recomienda, por ejemplo, el uso de los tradicionales sistemas de clasificación general del área de biblioteconomía y documentación, DDC, LCSH, los vocabularios de áreas científicas específicas, MEdical Subjects Heading, Tesauro de la NASA, la taxonomía ACM, grandes bases de datos y de conocimiento léxico, Cyc, WordNet, o los construidos por consorcios y organizaciones internacionales, tesauro de la UNESCO, o el de ARIADNE (figura 2.12).

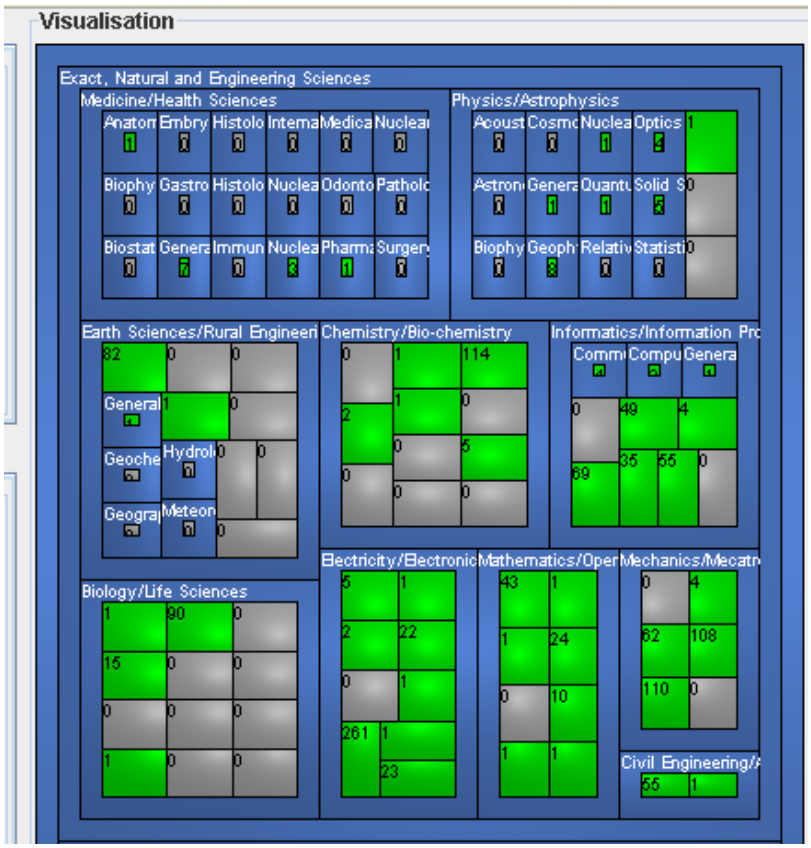


Figura 2.12. Presentación del tesauro ARIADNE del repositorio Knowledge Pool[59]. Los términos se estructuran en con la relación de generalización-especialización. Los números indican el número de recursos indexados con cada término[60]

[57] Lo que significa evaluados por algún comité u organización reconocida.
[58] Folksonomía (procede del inglés folksonomy) es un sistema de categorización que describe un dominio de objetos mediante etiquetas creadas colaborativamente; por ejemplo, del.icio.us (http://del.icio.us) categoriza enlaces favoritos, Flickr, fotos, tanzania o flof (http://flof.com.ar/), categoriza lugares.
[59] www.ariadne-eu.org/

Esta circunstancia en ciertos ámbitos, como la enseñanza en entornos virtuales, *e-learning*, constituye un serio inconveniente, porque los vocabularios "recomendados" para las colecciones de recursos educativos (Lee y Sugimoto, 2005) a) no se ajustan al dominio temático de las colecciones de objetos de contenido; b) no se ajustan al lenguaje de especialidad de los usuarios; y c) no se ajustan al propósito de uso docente e investigador. La solución es, por un lado, su adaptación al lenguaje y al dominio, pero esto es difícil de abordar para los usurarios porque requiere recursos técnicos y de personal. Por otro lado, aprender el vocabulario recomendado requiere un importante esfuerzo y tiempo por parte de los autores de los objetos de contenido, de los indexadores y de los usuarios de los repositorios de objetos de contenido. Además, esto no garantiza que sea capaz de describir con precisión la colección de recursos (Friesen, 2004; Heath et al., 2005; Hepp, 2007).

Los vocabularios controlados y metadatos para documentar los recursos educativos digitalizados tienen unas características propias. En primer lugar, su objetivo básico es la representación, búsqueda, navegación y selección de los recursos educativos digitalizados (ANSI/NISO Z39.19, 2005). La interoperabilidad es un objetivo secundario que, en la mayoría de los casos, se aborda con el uso de ontologías (Sampson et al., 2004). En segundo lugar, en vez de utilizar vocabularios generales recomendados, se construyen y utilizan vocabularios "propietarios" específicos de dominios y comunidades de usuarios particulares orientados a la enseñanza, al aprendizaje virtual, y a los dominios temáticos que abarcan los recursos educativos que se describen y clasifican[61] (Friesen, 2004; Lee y Sugimoto, 2005; CEN CWA 14871, 2003). En tercer lugar, es característico su uso didáctico como herramienta de representación conceptual del dominio de conocimiento al que pertenecen las colecciones de recursos para apoyar y reforzar el aprendizaje de los términos y conceptos del dominio (Greenberg et al., 2005; Sierra y Fernández-Valmayor, 2006). Finalmente, la cuarta característica es que estos vocabularios son construídos frecuentemente por grupos de profesores y estudiantes de forma inductiva y colaborativa, aprovechando el soporte de los campus virtuales (Dekkers y Feria, 2006;

[60] Fuente http://ariadne.cs.kuleuven.be/silo2006/visualbrowse.jsp

[61] Las razones, tal como hemos dicho antes, pueden estar en que los profesores, que son los autores, indexadores y usuarios de los recursos educativos que se indexan con vocabularios, necesitan invertir un esfuerzo grande en conocer los vocabularios que se consideran recomendados, sin la garantía de que representen adecuadamente las áreas específicas de conocimiento ni se ajusten a las necesidades docentes.

Sierra et al., 2005; Or-Bach, 2005), (también se pueden consultar numerosas experiencias en (Panizo et al., 2006).

En resumen, la experiencia actual de uso de metadatos y vocabularios para la explotación de colecciones de recursos educativos digitalizados indica que, en este contexto específico, la construcción y uso de vocabularios debería tener en cuenta la necesidad de: 1) ajustarse a los recursos e información de la colección; 2) ajustarse al lenguaje de especialidad de los profesores que son autores y usuarios de los recursos, 3) ajustarse a los propósitos didácticos, 4) encontrar mecanismos que faciliten, a los profesores, la construcción y la colaboración en la construcción de dichos recursos e información; 5) usar los vocabularios como marco conceptual de referencia del conocimiento y trabajo de cada comunidad docente-discente específica; y 6) usar los vocabularios como un recurso educativo para aprender y comprender el lenguaje de especialidad y el conocimiento que contienen. Los entornos, modelos y metodologías de construcción y uso de vocabularios serán más eficaces para la enseñanza-aprendizaje si incorporan estas necesidades a sus objetivos de diseño.

2.6 Tipos de vocabularios para la explotación de recursos didácticos digitalizados

Esta sección proporciona una descripción de la tipología de vocabularios orientados a la enseñanza y aprendizaje en entornos virtuales, *e-learning*, que es imprescindible para poder seleccionar el más adecuado a cada situación[62]. Está basada en los estándares de construcción de vocabularios monolingües (representados por la última versión del ANSI/NISO Z39.19 del año 2005) y en las experiencias de uso en *e-learning* (CEN CWA 14871, 2003). Se distinguen:

- Vocabularios simples o listas de valores;
- Clasificaciones y taxonomías;
- Tesauros;
- Ontologías;
- Diccionarios y glosarios.

[62] La selección de una opción depende de diversos factores: i) la funcionalidad para la que se aplica, ii) los usuarios a los que se dirige, iii) la compatibilidad con el sistema de información donde se integra, iv) los recursos (económicos, personal y temporales) para su construcción y mantenimiento, v) la cobertura, vi) el tamaño, vii) la naturaleza de los objetos de contenido y viii) el tipo de consultas, búsquedas y perfiles de usuario (Aitchison et al., 2000).

2.6.1. Vocabulario simple o lista de valores

Una lista de términos es la forma más simple de un vocabulario[63]. Los términos son el único componente y, aunque pueden ir acompañados de identificadores únicos de tipo alfanumérico, no contienen definiciones u otra información asociada. La estructura es plana: lineal o jerárquica, con dos o tres niveles máximos y, normalmente, ordenada alfabéticamente.

Los vocabularios simples son apropiados para la indexación y búsqueda basada en atributos o propiedades. Por ejemplo, en las descripciones con metadatos, estos vocabularios definen un conjunto finito de valores de referencia para una propiedad determinada (figura 2.13). Desde el punto de vista informático, son muy eficientes y facilitan la interoperabilidad, pero su capacidad de representación semántica está limitada a descripciones generales poco refinadas, lo que podríamos considerar granularidad gruesa.

Semantics	Science Type	Exact, Natural and Engineering Sciences
	Main Discipline	Informatics/Information Processing
	Sub-Discipline	Computer-Human Interaction
	Main Concept	x ariadne
Educational	Intended End User Role	Learner
	Document type	Expositive
	Document Format	Slides
	Typical Learning Time (minutes)	20
Technical	Main File Name	20040916_ICETA_Kosice_video.pdf
	Media (MIME) Type	application/pdf
	Operating System (OS)	Multi-OS

Figura 2.13. Extracto de los metadatos LOM de un recurso educativo del repositorio europeo ARIADNE. Los términos utilizados como valores de los subelementos del elemento Educational provienen de vocabularios simples (excepto Time). También los términos de los campos Media (MIME) Type y Operating System (OS)

La figura 2.14 muestra un ejemplo típico de uso de un vocabulario simple, el código de lenguas ISO 3166, para dar valores al atributo *language*. La figura 2.13 muestra los

[63] Se corresponde con el tipo de datos enumerado de los lenguajes de programación.

metadatos de un recurso educativo del repositorio europeo ARIADNE. Esta ficha de metadatos contiene varios elementos o atributos de LOM con valores tomados de vocabularios simples como el atributo técnico *media (MIME) type* asociado al vocabulario estándar RFC 1521.

```
<string language="de">Allgemein/Verschiedenes</string>
<string language="en">General/Sundry</string>
<string language="es">Generalidades/Varios</string>
<string language="fr">G&eacute;n&eacute;ralit&eacute;s/Divers</string>
<string language="it">Generalit&agrave;/Varie</string>
<string language="nl">Algemeen/Diversen</string>
<string language="ro">Generalitati/Diverse</string>
```

Figura 2.14. Uso del vocabulario estándar ISO 3166 para el lenguaje (de, en, es,etc.)

2.6.2. Clasificaciones y taxonomías

La diferencia entre clasificación y taxonomía es muy sutil[64]. En primer lugar, definimos el concepto de clasificación. Una clasificación es un vocabulario formado por categorías que pueden contener términos. Las categorías se presentan, normalmente, ordenadas alfabéticamente (figura 2.15), o jerárquicamente, si entre ellas se relacionan por generalización-especialización (figura 2.16.). Los términos, si existen, están incluidos en las categorías (figura 2.17).

Figura 2.15. Clasificación alfabética de artículos, con indicación de la frecuencia del número de accesos a cada término mediante su mayor o menor tamaño[65]

[64] De hecho, en práctica se utilizan frecuentemente los dos términos como sinónimos, por ejemplo, la figura 2.16 muestra una parte de la taxonomía de la ACM para Ciencias de la Computación que, sin embargo, se denomina "sistema de clasificación".

[65] Fuente: http://cent.uji.es/octeto/taxonomy/

A. General Literature
B. Hardware
C. Computer Systems Organization
D. Software
E. Data
F. Theory of Computation
G. Mathematics of Computing
H. Information Systems
H.0 GENERAL
H.1 MODELS AND PRINCIPLES
H.2 DATABASE MANAGEMENT (E.5)
H.3 INFORMATION STORAGE AND RETRIEVAL
H.4 INFORMATION SYSTEMS APPLICATIONS
H.5 INFORMATION INTERFACES AND PRESENTATION (e.g., HCI) (I.7)
H.m MISCELLANEOUS
I. Computing Methodologies
J. Computer Applications
K. Computing Milieux

Figura 2.16. Nivel superior y siguiente en categoría H del Sistema de Clasificación de la ACM[66]

8 Language. Linguistics. Literature

81 Linguistics and language

81'1/'4...	Facets of linguistics
81'1	General linguistics
81'2...	Semiotics. Psycholinguistics. Sociolinguistics. Usage. Dialectology
81'3...	Mathematical and applied linguistics. Phonetics. Graphemics. Grammar. Semantics
81'4...	Text linguistics. Discourse analysis. Typological linguistics
811	**Languages**
811.1/.8	Individual (natural) languages Parallel with Table 1c - Languages
811.9	Artificial languages

Figura 2.17. Categoría raíz número 8 "Lengua, Lingüística,Literatura" de la Clasificación Universal Decimal con algunos de los términos que incluye[67]

Una taxonomía es una clasificación de una especialidad, disciplina o área temática particular (CEN CWA 14871, 2003). Es muy frecuente que estén integradas en vocabularios más generales, por ejemplo, ontologías, constituyendo las representaciones del conocimiento de los dominios específicos[68].

La estructura de una taxonomía es una jerarquía. En el nivel más alto se sitúan los términos o categorías generales. Los sucesivos niveles de la jerarquía refinan los

[66] Association for Computing Machinery: http://www.acm.org/class/1998/homepage.html
[67] Fuente: http://www.udcc.org/index.htm
[68] La ontología Cyc, por ejemplo, contiene 9 taxonomías http://taxonomies.cyc.com/

términos o categorías del nivel superior. Los descriptores, además, pueden tener un identificador alfanumérico (figura 2.16).

La aplicación de clasificaciones y taxonomías para describir conceptualmente los recursos didácticos digitalizados consiste en asociar cada recurso a las categorías adecuadas. La descripción semántica del recurso es un conjunto de caminos, llamados caminos taxonómicos, desde las categorías terminales denominadas "hoja" que los incluyen hasta las categorías "raíz" de la clasificación o taxonomía (figura 2.18). Las herramientas de indexación y búsqueda por conceptos y temas de los repositorios de recursos educativos están basadas, en su mayoría, en categorías y taxonomías (figura 2.19).

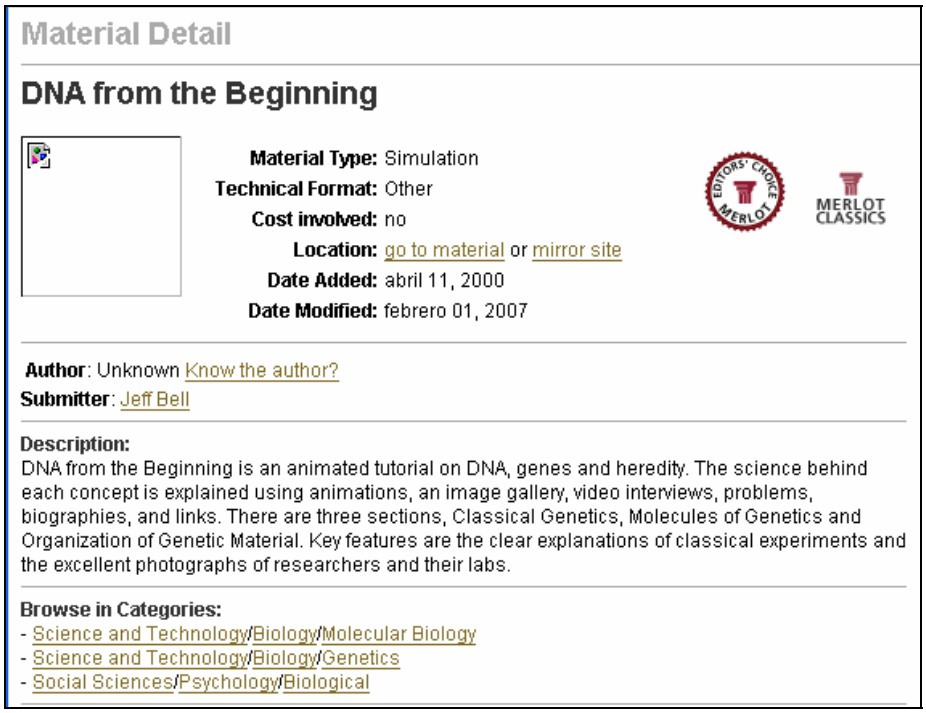


Figura 2.18. Clasificación de un recurso con varios caminos taxonómicos (Browse in Categories)

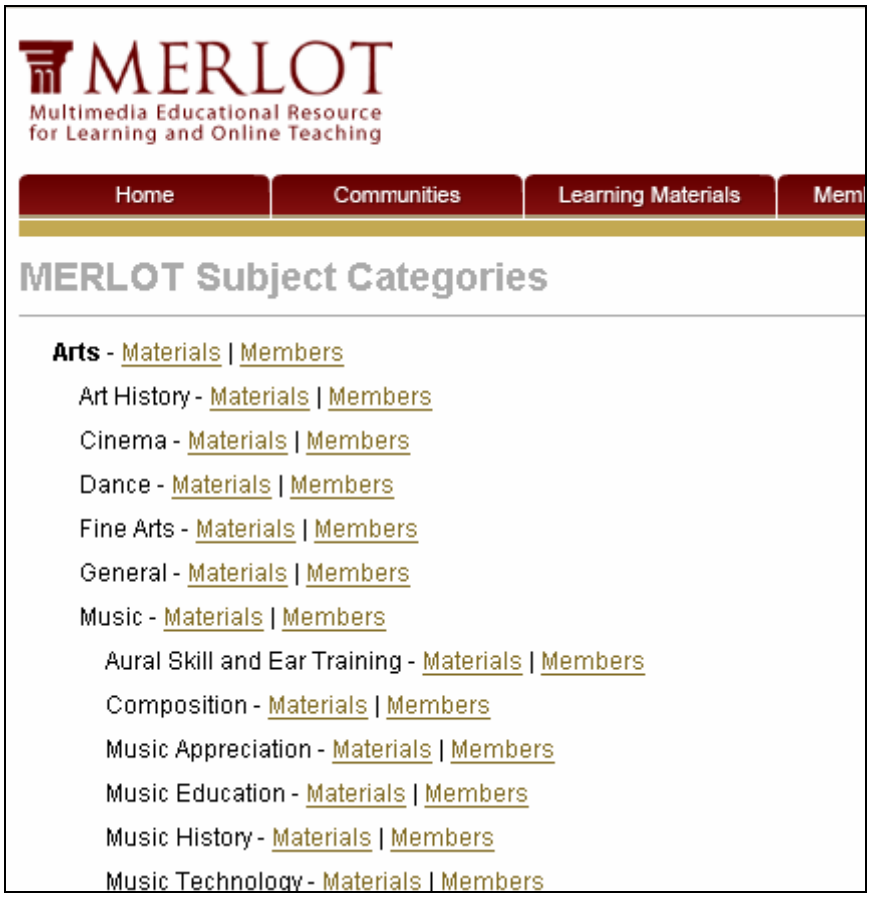


Figura 2.19. Categoría/taxonomía del repositorio de recursos educativos MERLOT

2.6.3. Tesauros

Un tesauro es un vocabulario limitado, generalmente de palabras especializadas, dotado de sus correspondencias semánticas, y elegido para que represente las nociones que figuran en un contexto dado para su empleo en informática y en el establecimiento de índices (Martínez de Sousa, 1995). En Aitchison et al. (2000), un tesauro es "el vocabulario de un lenguaje controlado de indexación, formalmente organizado de forma que se hacen explícitas a priori las relaciones entre conceptos". Por su parte, el estándar ANSI/NISO Z39.19 (2005) define el tesauro como "un vocabulario controlado organizado y estructurado de forma conocida y donde las relaciones entre términos de equivalencia, homógrafos, jerárquicas y asociativas se visualizan claramente mediante marcadores estándares y recíprocos"[69].

La diferencia con otros tipos de vocabularios está en que su prioridad es representar formalmente las relaciones semánticas entre conceptos utilizando términos del lenguaje natural. Los vocabularios simples no contienen relaciones semánticas, las

[69] Resulta curioso comprobar cómo se ha especializado el concepto de tesauro con los nuevos usos en la Tecnología Lingüística.

clasificaciones y taxonomías explicitan sólo la relación de generalización/especialización. Los diccionarios y glosarios, como veremos, contienen descripciones de los conceptos en lenguaje natural y las relaciones entre conceptos no son siempre explícitas (figura 2.20). Las ontologías representan los conceptos y sus relaciones a un nivel de abstracción mayor, sin bajar al lenguaje. Representan los significados extraídos del significante.

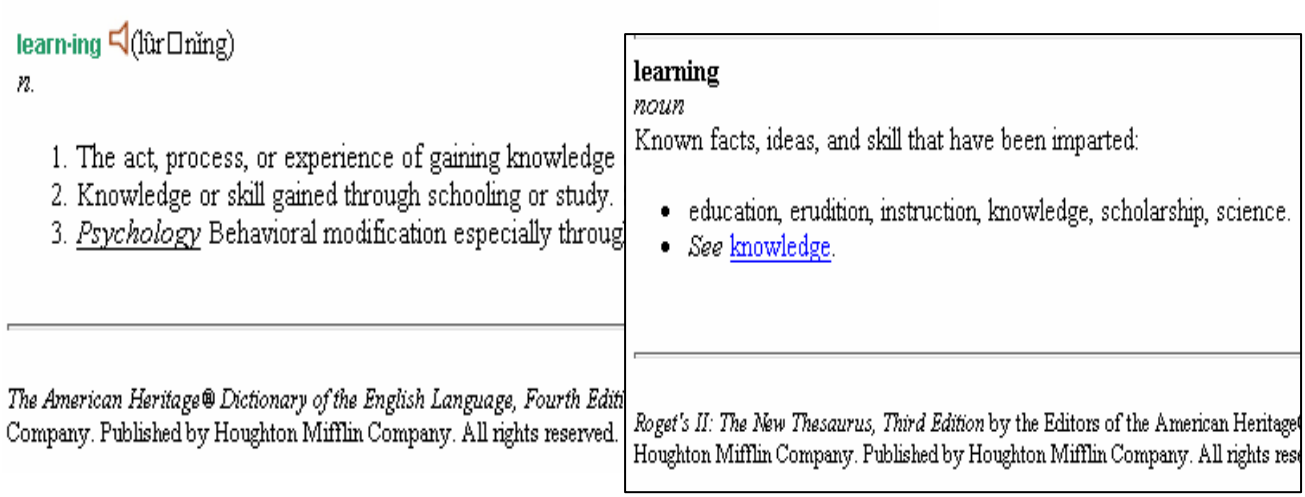


Figura 2.20. Contenido y estructura de la entrada learning en el diccionario electrónico The American Heritage (izquierda) y en el Tesauro de Roget (derecha)

La naturaleza altamente relacional del tesauro comporta una estructura compleja que necesita la definición de un esquema de datos formal para mantener la consistencia durante su construcción y actualización. Este esquema define las relaciones semánticas[70], los términos, las categorías y la información no relacional, si existe, que va a contener el tesauro. En el siguiente capítulo se revisarán los modelos de datos utilizados para construir los esquemas de los tesauros de explotación y las recomendaciones estándar para su presentación según el tipo de utilización (figuras 2.21, 2.22 y 2.23)[71].

[70] Las relaciones estándar de asociatividad, equivalencia y jerarquía siempre pueden extenderse, con el objeto de poder refinar las descripciones conceptuales y obtener mejores resultados en las búsquedas (Tudhope et al., 2001).

[71] Se recomiendan cinco tipos de presentación: alfabética, lista permutada, jerárquica, término seleccionado y gráfica.

ACEPTABILIDAD (LINGÜISTICA)
T.E.:
AGRAMATISMO
GRAMATICALIDAD (LINGÜISTICA)
T.G.:
SOCIOLINGÜISTICA
ACTOS DE HABLA (LINGÜISTICA)
T.R.:
LENGUAJE Y LENGUAS-FILOSOFIA
ACTUACION Y COMPETENCIA (LINGÜISTICA)
U.S.:
COMPETENCIA Y ACTUACION (LINGÜISTICA)
ANAFORA (LINGÜISTICA)
N.A.:
V. A. EL SUBENC.-ANAFORA BAJO LENGUAS Y GRUPOS DE LENGUAS
T.R.:
GRAMATICA COMPARADA Y GENERAL
U.P.:
REFERENCIAS CRUZADAS (LINGÜISTICA)

Figura 2.21. Extracto del tesauro de la Biblioteca de la UCM para la gestión documental. Presentación alfabética estándar

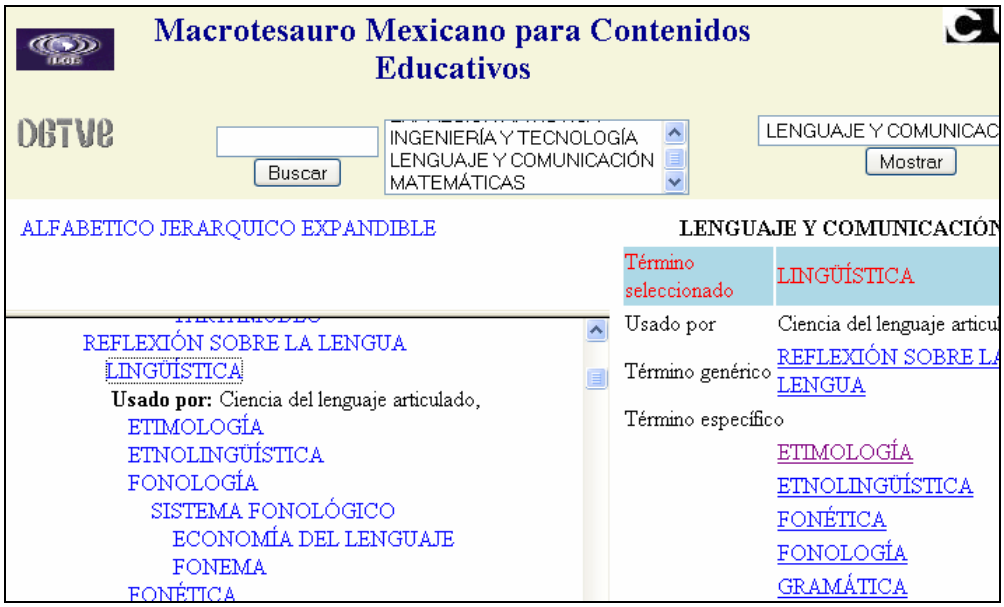


Figura 2.22. Presentación jerárquica (izquierda) y por término detallado (derecha) del tesauro MMCE para la descripción de contenidos educativos[72]

[72] Fuente: http://cuib.unam.mx/~tesauro

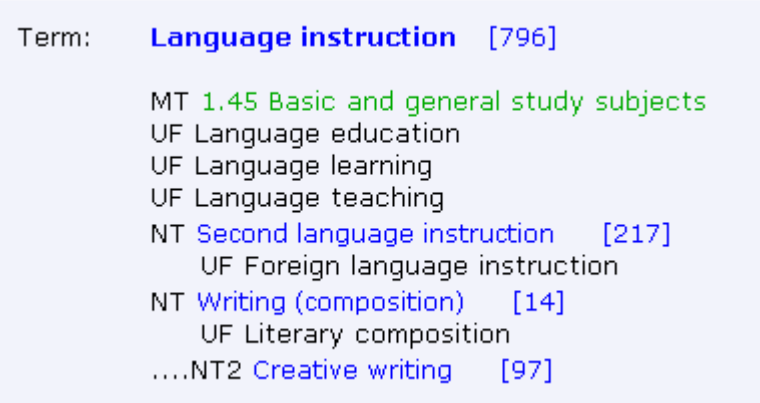


Figura 2.23. Presentación por término seleccionado del tesauro del repositorio de recursos documentales de la UNESCO[73]*. A la derecha de los términos se indica y enlaza al número de recursos asociados al término.*

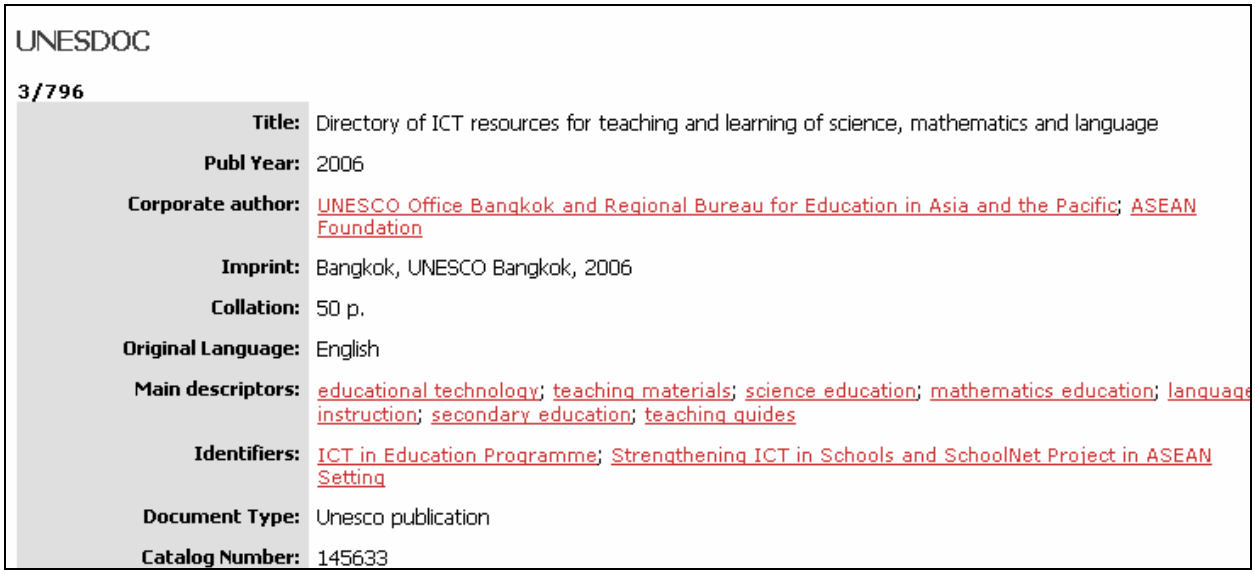
Una de las principales aplicaciones de los tesauros es, actualmente, la RI, pero no es la única. Otros usos incluyen el apoyo en la comprensión general de un área de conocimiento proporcionando mapas conceptuales y esquemas conceptuales que muestran las interrelaciones entre los conceptos y las entidades, la búsqueda de términos alternativos durante la escritura o lectura de textos, el aprendizaje de los términos de una disciplina o la generación de listas de palabras clave (Aitchison et al., 2000).

Los tesauros en RI se aplican a la exploración, búsqueda y selección de información y recursos en espacios de almacenamiento muy grandes como la Web, bases de datos documentales o repositorios. Uno de los ámbitos de aplicación es el *e-learning*, especialmente para la gestión de contenidos en los CMS o LCMS, que describiremos con mayor detalle en el próximo capítulo (figuras 2.22, 2.23, 2.24 y 2.25). En estos sistemas, la indexación, que es por asignación, suele utilizar principalmente los tesauros, aunque también las clasificaciones y taxonomías. Los contenidos y recursos se asocian a una o varios términos del tesauro (figura 2.23) y durante la búsqueda se aplica, normalmente, una estrategia de postcoordinación en la que se calculan todas las combinaciones de los términos de la consulta que estén relacionados en el tesauro para precisar y completar la consulta, reduciendo, de este modo, el ruido o resultados no deseados, o buscando alternativas cercanas al silencio, es decir, ningún resultado (figura 2.25) (CEN CWA 14871, 2003).

La descripción conceptual de los recursos con tesauros se realiza asociando el recurso al conjunto de términos interrelacionados, los términos de indexación, del tesauro que

[73] Fuente: http://databases.unesco.org/thesaurus/

mejor describen su contenido. Se puede decir que el recurso se coloca en el tesauro. Para simplificar la descripción, ésta se forma sólo con el subconjunto de términos con los que está directamente asociado el contenido del recurso (figura 2.24). Los algoritmos de búsqueda o las presentaciones gráficas se encargan de obtener del tesauro el resto de los términos con los que, el recurso, está indirectamente relacionado y la naturaleza de éstas relaciones (figura 2.26).


UNESDOC

3/796

Title:	Directory of ICT resources for teaching and learning of science, mathematics and language
Publ Year:	2006
Corporate author:	UNESCO Office Bangkok and Regional Bureau for Education in Asia and the Pacific; ASEAN Foundation
Imprint:	Bangkok, UNESCO Bangkok, 2006
Collation:	50 p.
Original Language:	English
Main descriptors:	educational technology; teaching materials; science education; mathematics education; language instruction; secondary education; teaching guides
Identifiers:	ICT in Education Programme; Strengthening ICT in Schools and SchoolNet Project in ASEAN Setting
Document Type:	Unesco publication
Catalog Number:	145633

Figura 2.24. El campo "Main descriptors" contiene una descripción semántica del recurso formada simplemente por el conjunto de términos del tesauro a los que está asociado.

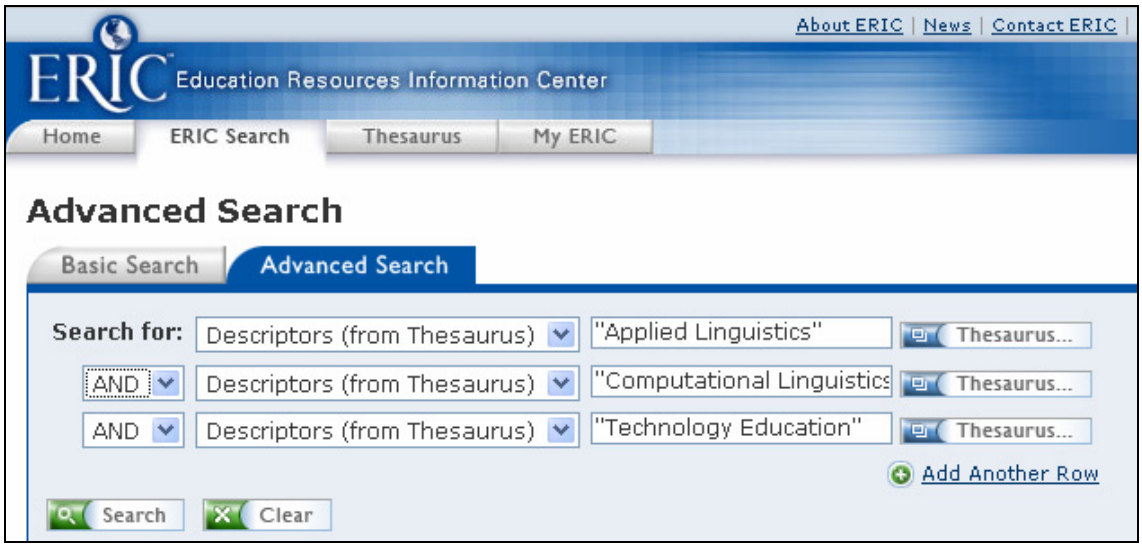


Figura *2.25. Búsqueda con ayuda del tesauro (ERIC Thesauro)*[74]

[74] Fuente: http://www.eric.ed.gov/

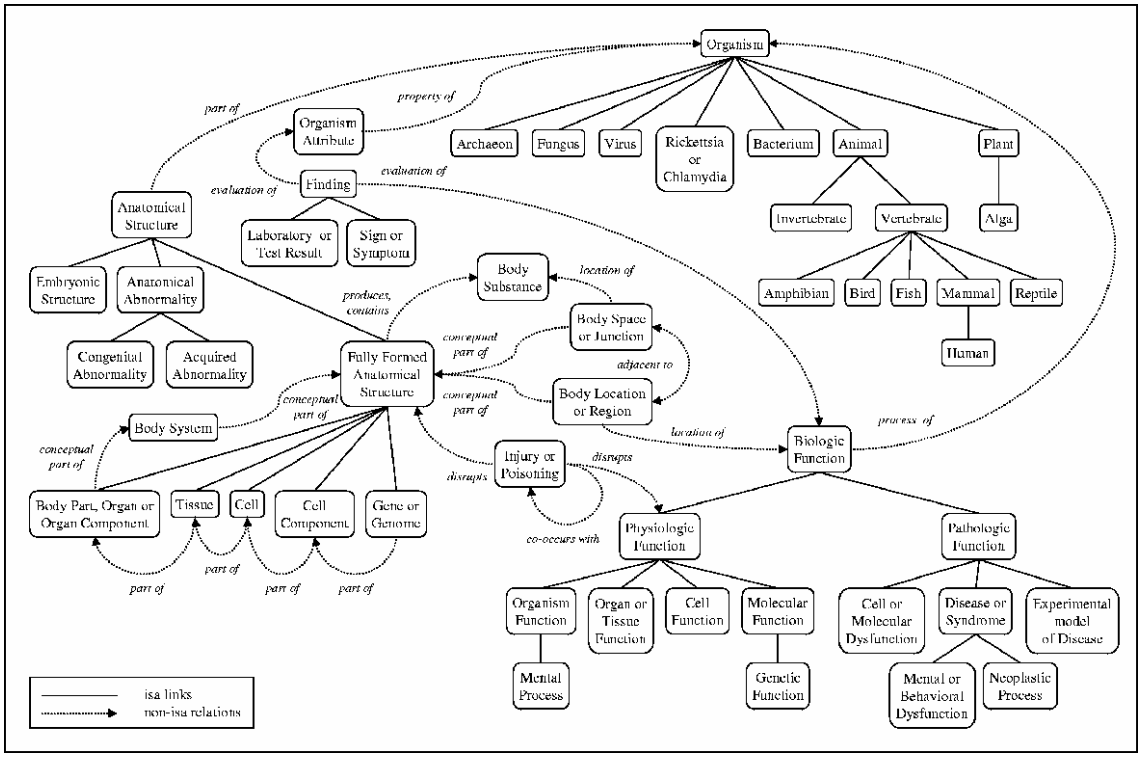


Figura 2.26. Presentación gráfica de una parte de la Red Semántica del tesauro UMLS[75]

A partir de estas representaciones conceptuales, las colecciones de recursos pueden explorarse utilizando el tesauro como un esquema de organización conceptual sobre el que construir sistemas de navegación. En este caso, el tesauro es un sistema de representación del contenido y el significado de los recursos que tienen indexados, ampliando las posibilidades de búsqueda y navegación sobre un tema o materia o materias relacionadas[76] (figura 2.27).

[75] Unified Medical Language System (UMLS, 2009).

[76] En la navegación el usuario puede ir descartando interactivamente aquellos términos del tesauro del repositorio que no se ajusten a su petición con el objetivo de ir precisando los resultados. También puede ampliar los resultados buscando nuevos recursos a partir de los términos relacionados con el término o términos de búsqueda.

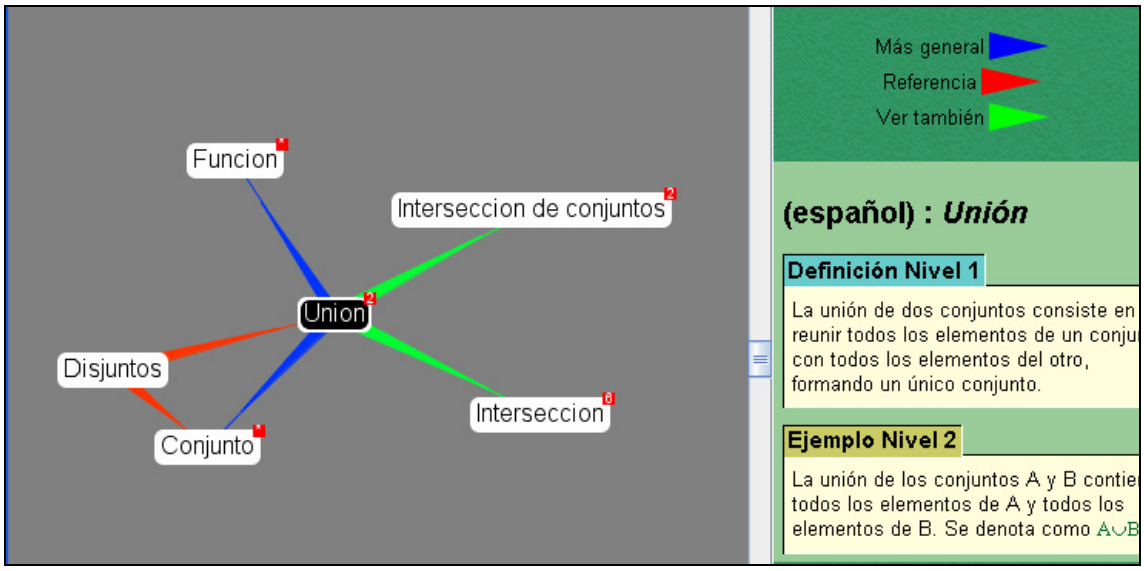


Figura 2.27. Presentación gráfica del tesauro thesaurus.maths.org con la organización terminológica y conceptual de la disciplina matemáticas

Cara a la representación y recuperación de objetos de contenido, los tesauros tienen la ventaja de de ser capaces de representar conceptos y relaciones entre conceptos utilizando lenguaje natural o de especialidad, pero sin la ambigüedad intrínseca del lenguaje natural. Sin embargo, esto también es una limitación en cuanto a su poder expresivo. Los tesauros no son una representación conceptual del dominio de conocimiento ni tampoco son una representación en lenguaje natural, se trata de un sistema puente entre las representaciones conceptuales y el lenguaje.

2.6.4. Ontologías

De todos los tipos de vocabularios, las ontologías son las representaciones que se pueden considerar conceptuales. Desde el punto de vista de los sistemas de información, la ontología se define como la especificación de una conceptualización (Gruber, 1993), o, de forma más detallada, *la especificación formal de una conceptualización compartida* (Borst, et al., 1997). Una *conceptualización* es una vista abstracta y simplificada del mundo que se desea representar –modelo del dominio-; por *compartida* se entiende su carácter consensuado, y *formal* significa que es explícita y precisa por lo que es posible procesarla automáticamente (figura 2.28). El término ontología proviene de la Filosofía, pero se utiliza –sobre todo desde la década de los noventa- ampliamente en Inteligencia Artificial, Lingüística Computacional, Tecnología Educativa y en Biblioteconomía y Documentación para tratar la representación, gestión e intercambio del conocimiento. Las ontologías aportan una compresión compartida y consensuada de

un dominio del conocimiento, que puede ser comunicada entre personas y sistemas heterogéneos (Lamarca, 2007).

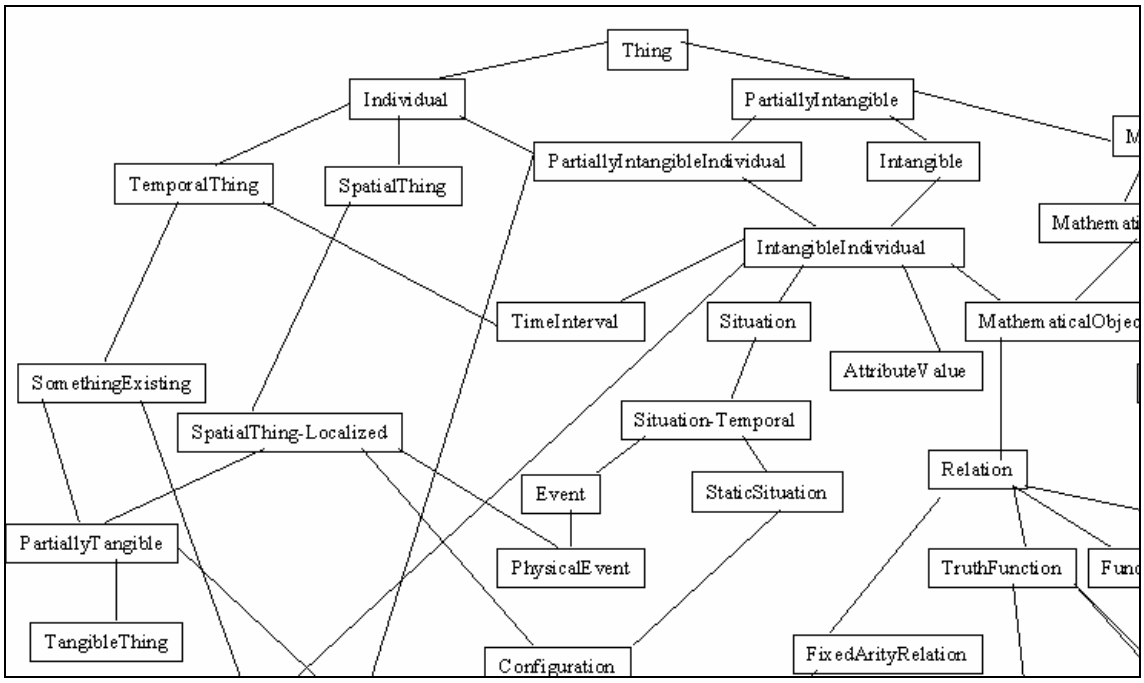


Figura 2.28 Vista parcial de la ontología Cyc (parte superior ó "upper ontology"[77]

El objetivo de las ontologías es facilitar la compartición y comprensión de un dominio de conocimiento entre personas o sistemas o ambos, proporcionando (1) una forma de representar y compartir el conocimiento común, (2) un formato de intercambio de conocimiento, (3) un protocolo específico de comunicación, y (4) la posibilidad de reutilizar conocimiento de distintas fuentes (Berners et al., 2001).

Las ontologías están formadas por el conjunto de entidades, relaciones y conceptos que definen un dominio y utilizan un vocabulario y semántica comunes (figura 2.28). Es un modelo compartido para interpretar conceptos y ejemplares o instancias de los conceptos (figura 2.29).

[77] Fuente: http://www.cyc.com/cycdoc/upperont-diagram.html

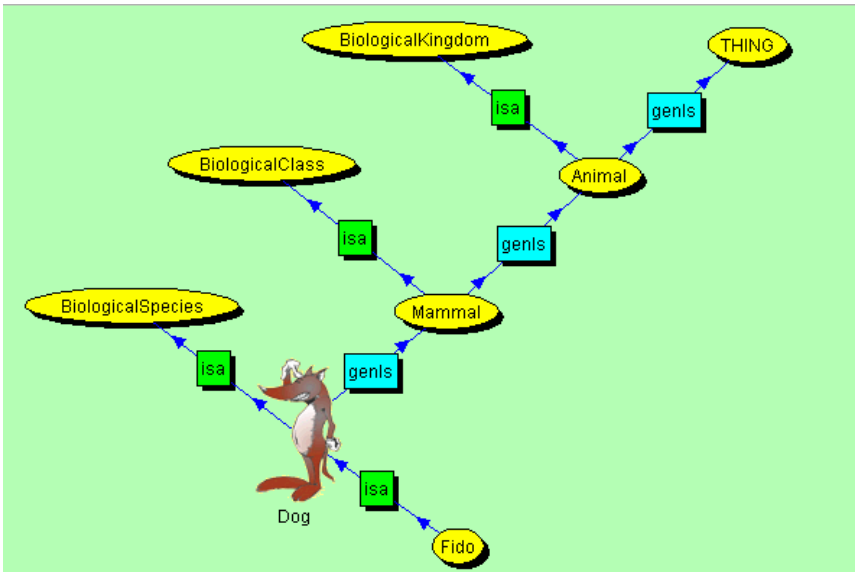


Figura 2.29 Definición de la instancia "Fido" en la ontología Cyc

El modelo de ontología es muy variado y depende del tipo, propósito y dominio de la ontología (Van Heijst et al., 1997). Existen múltiples definiciones de tipos de ontologías, las más utilizadas pueden consultarse en Sowa (2000), Fensel et al. (2001), McGuiness (2003). En Sowa (2000) se distinguen entre ontologías formales, terminológicas, prototipos y mixtas. En las *ontologías formales* los conceptos se definen mediante axiomas y definiciones en lógica o en algún lenguaje de programación que pueda traducirse automáticamente a lógica (figura 2.30). Estas ontologías son las más potentes en cuanto que permiten la consulta del conocimiento explícitamente almacenado, como el conocimiento implícito que se obtiene por inferencias y razonamientos automáticos, pero son costosas de construir y de mantener.

Abstract (A).

- Ninguna abstracción τιενε υνα λοχαλιζαχι ſν εσπαχιαλ: ~(∃*x*:Abstract)(∃*y*:Place)loc(*x,y*).
- Ninguna abstracción aparece en un momento temporal: ~(∃*x*:Abstract)(∃*t*:Time)pTim(*x,t*).

Figura 2.30. Definición del concepto "Abstracción" en la ontología KR Ontology (Sowa, 2000)

Las *ontologías terminológicas* expresan los conceptos con términos del lenguaje natural y con relaciones semánticas (Sowa, 2000). Un ejemplo es WordNet, donde los conceptos están asociados a uno o varios términos del lenguaje y están relacionados

mediante relaciones semánticas que dependen de la categoría gramatical del término; en los sustantivos son de hiponimia-hiperonimia y parte-todo. Normalmente las *ontologías superiores*, que capturan el conocimiento en general sobre el mundo y proporcionan nociones básicas y conceptos abstractos, son ontologías formales, mientras que las ontologías más específicas, correspondientes a dominios o áreas de conocimiento determinados, son ontologías terminológicas[78].

Las *ontologías prototipo* definen las categorías mediante sus instancias: para cada categoría c de la ontología debe existir un prototipo o instancia, p, y una función distancia semántica, d(x,y,c), que mide la similitud o diferencia entre cualquier par de instancias x, y de c. Las nuevas instancias se clasifican en categorías respecto a sus similitudes con las instancias de cada categoría.

Finalmente, las *ontologías mixtas* contienen categorías definidas formalmente, normalmente las categorías superiores o más abstractas, y categorías definidas mediante prototipo.

En cualquiera de los tipos, la estructura de las ontologías es compleja: constituyen estructuras matemáticas organizadas con relaciones de orden parcial, como *subtipo-tipo* –llamada hiponimia-hiperonimia en lexicografía-, o *parte-todo* –meronimia. La estructura que resulta se llama en matemáticas retículo y en ella pueden distinguirse uno o varios tipos de jerarquías. Las operaciones y propiedades definidas en los retículos pueden programarse para gestionar automáticamente el conocimiento que contienen.

En el contexto del *e-learning*, una de las aplicaciones de las ontologías es la descripción completa de un sistema de enseñanza-aprendizaje virtual -tipos de usuarios, la organización y administración académica, cursos, actividades, recursos, etc.- con diversos fines, como personalizar y adaptar la enseñanza a las necesidades de cada estudiante -Tutores Inteligentes, facilitar el aprendizaje colaborativo en un entorno distribuido, facilitar el intercambio y reutilización de la información y conocimiento entre los sistemas y personas, y, finalmente, facilitar la gestión y acceso de los repositorios de recursos didácticos digitalizados (CEN CWA 14871, 2003; Staab y Studer, 2004; Sampson et al., 2004; Méndez, 2005; Dekkers y Feria, 2006; Panizo et al., 2006).

Respecto al papel que juegan las ontologías en la gestión y acceso a los repositorios de recursos educativos, éstas se utilizan como índices para localizar términos en búsquedas

[78] Taxonomías en Cyc: http://www.cyc.com/cyc/products/taxonomies

basadas en conceptos (Hernández y Saiz, 2007). También se utilizan como mapas y esquemas conceptuales del dominio, al igual que los tesauros, sobre los que el profesor y el estudiante pueden navegar para localizar recursos con el contenido, tópico o conceptos relacionados (ver por ejemplo Jovanović et al., 2005) (figura 2.31).

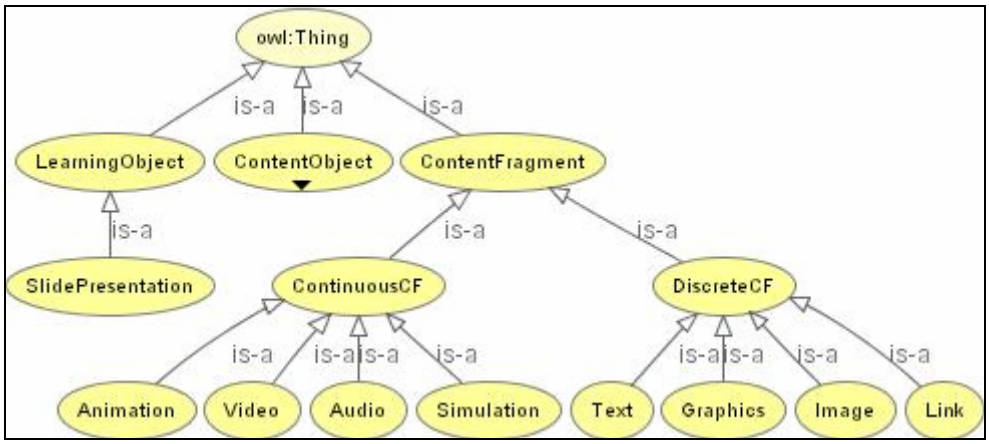


Figura 2.31 Conceptos del nivel superior de la ontología ALOCoM[79]

En la práctica, la distinción entre ontologías y tesauros no está siempre clara. En la bibliografía citada a lo largo de esta sección, encontramos numerosos ejemplos de tesauros denominados ontologías, especialmente cuando se trata de ontologías terminológicas que modelan dominios concretos de conocimiento. Por ejemplo, en la clasificación de ontologías de McGuiness (2003) se incluye los vocabularios o lexicones -listas simples, glosarios y tesauros- como tipos de ontologías.

Sin embargo en Sowa (2000b) y Hirst (2004) se establecen con claridad las diferencias entre los vocabularios entendidos como lexicones –listas de valores, taxonomías, tesauros, diccionarios y glosarios- y las ontologías: (1) las ontologías constituyen las representaciones conceptuales del dominio de conocimiento a un nivel semántico profundo; (2) deben ser representaciones independientes de las lenguas naturales (aunque sea cercana para ser legible)[80]; y (3) pueden incluir conocimiento inferencial (reglas) que permitan razonar a partir del conocimiento explícitamente almacenado. J. F. Sowa (2005) definió acertadamente esta diferencia con la siguiente frase: "el lexicón es el puente entre el lenguaje y el conocimiento que se expresa con ese lenguaje".

[79] ALOCoM es una ontología que define un Modelo de Contenido para los Objetos de Aprendizaje del repositorio europeo ARIADNE (Ariadne Learning Objects Content Model) (Knight et al., 2005).

[80] Por esta razón, se han utilizado, con un éxito relativo, como módulos interlingua en sistemas multilingües (Carbonell et al., 1992).

En cualquier caso, las ontologías se prefieren para facilitar la interoperabilidad entre agentes, aplicaciones y personas, mientras que los tesauros se prefieren como mecanismos de indexación, búsqueda y navegación de información y de colecciones de recursos digitalizados.

2.6.5 Glosarios y diccionarios

Los glosarios y diccionarios son vocabularios organizados y escritos para uso humano aunque se presenten en formato electrónico, entre otras razones, porque la información que contienen no siempre es explícita, está semi-estructurada[81] y su interpretación depende de la competencia lingüística y conocimiento de mundo del usuario (figura 2.32) (Fernández-Pampillón y Matesanz, 2003). En las tablas 2.1 y 2.2 de la sección precedente 2.2, se recogen algunas de las definiciones de diccionario y glosario. La diferencia entre ambos puede establecerse por su cobertura, complejidad y propósito. En este sentido los glosarios tienen un número limitado de entradas y un contenido más específico que los diccionarios; su propósito es definir con precisión los términos específicos de un texto u obra (figura 2.33) o de una materia determinada (figura 2.34) para ayudar al lector en su comprensión (Blustein y Noor, 2004).

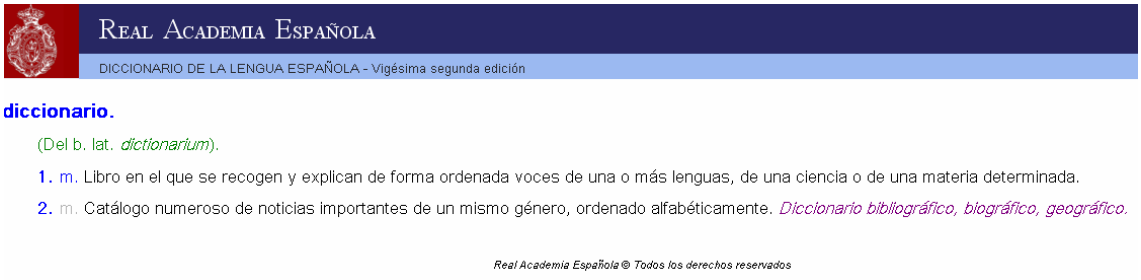
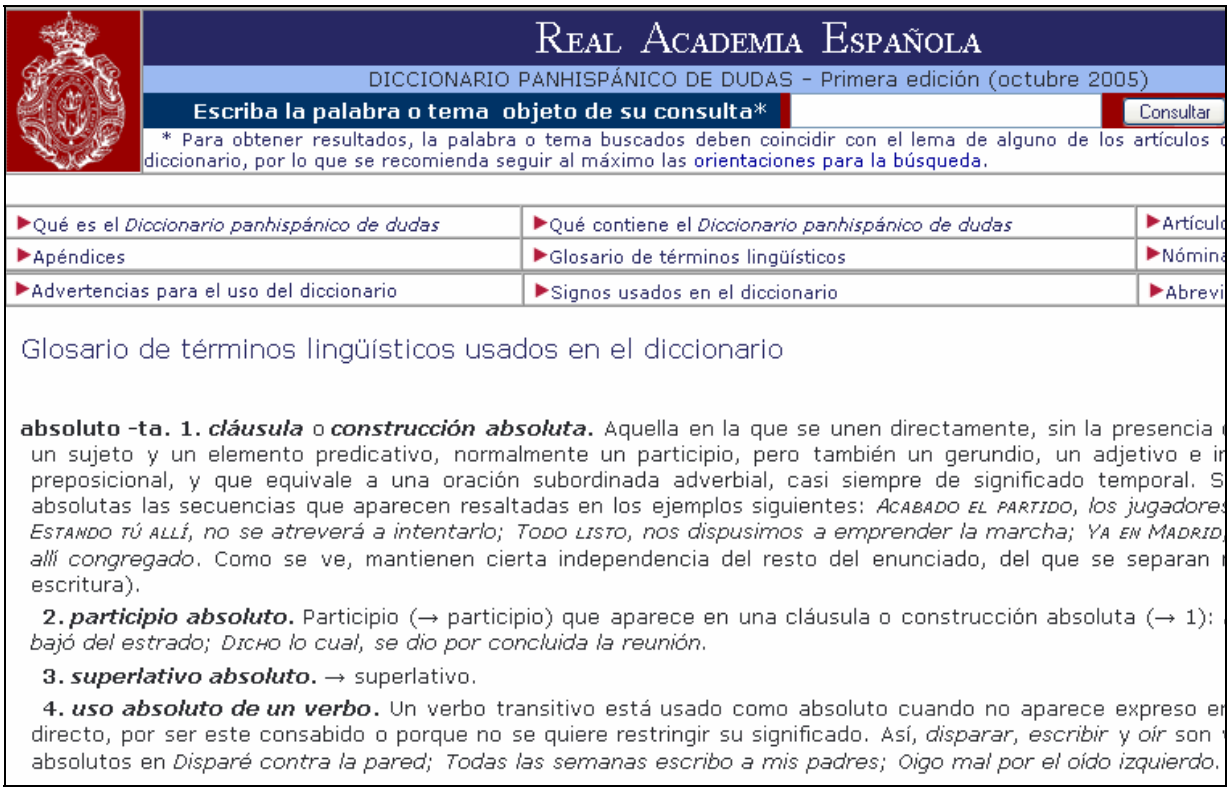


Figura 2.32. Consulta de la palabra "diccionario" en el (DRAE, 2001) en su versión en línea.

[81] La información está estructurada cuando se dispone de un esquema de organización de la información independiente del contenido de la información. Por ejemplo los documentos XML que disponen de una definición estructural con una DTD o un XML Schema. En el caso de información semi-estructurada la estructura está embebida en el contenido y se describe directamente utilizando una sintaxis simple (Abiteboul, et al., 2000). Así, los diccionarios tienen definidos unas convenciones de presentación de la información que formatean y distinguen cada elemento informacional. Por ejemplo, el lema […]"Encabeza el artículo y aparece representado en letra negrita, ya sea redonda (p. ej., bizcocho) o cursiva (boutique). Este segundo tipo de representación se reserva, como antes veíamos (§ 2.5), para los extranjerismos no adaptados en su pronunciación u ortografía a las reglas generales del español." [...](DRAE, 2001).



REAL ACADEMIA ESPAÑOLA

DICCIONARIO PANHISPÁNICO DE DUDAS - Primera edición (octubre 2005)

Escriba la palabra o tema objeto de su consulta* Consultar

* Para obtener resultados, la palabra o tema buscados deben coincidir con el lema de alguno de los artículos d diccionario, por lo que se recomienda seguir al máximo las orientaciones para la búsqueda.

▶Qué es el *Diccionario panhispánico de dudas* ▶Qué contiene el *Diccionario panhispánico de dudas* ▶Artículo

▶Apéndices ▶Glosario de términos lingüísticos ▶Nómina

▶Advertencias para el uso del diccionario ▶Signos usados en el diccionario ▶Abrevi

Glosario de términos lingüísticos usados en el diccionario

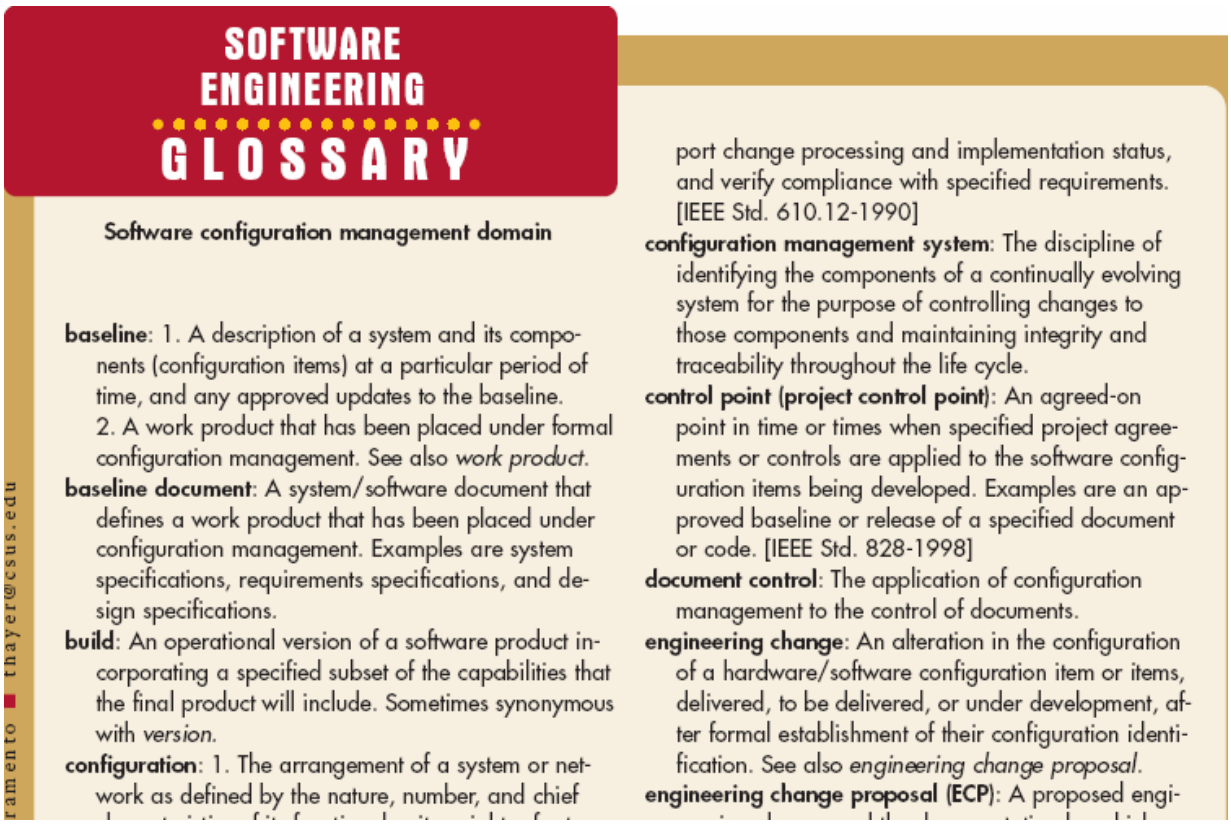
absoluto -ta. 1. ***cláusula*** o ***construcción absoluta.*** Aquella en la que se unen directamente, sin la presencia un sujeto y un elemento predicativo, normalmente un participio, pero también un gerundio, un adjetivo e ir preposicional, y que equivale a una oración subordinada adverbial, casi siempre de significado temporal. S absolutas las secuencias que aparecen resaltadas en los ejemplos siguientes: *Acabado el partido, los jugadores Estando tú allí, no se atreverá a intentarlo; Todo listo, nos dispusimos a emprender la marcha; Ya en Madrid, allí congregado*. Como se ve, mantienen cierta independencia del resto del enunciado, del que se separan escritura).

2. ***participio absoluto.*** Participio (→ participio) que aparece en una cláusula o construcción absoluta (→ 1): *bajó del estrado; Dicho lo cual, se dio por concluida la reunión.*

3. ***superlativo absoluto.*** → superlativo.

4. ***uso absoluto de un verbo.*** Un verbo transitivo está usado como absoluto cuando no aparece expreso en directo, por ser este consabido o porque no se quiere restringir su significado. Así, *disparar, escribir* y *oír* son absolutos en *Disparé contra la pared; Todas las semanas escribo a mis padres; Oigo mal por el oído izquierdo.*

Figura 2.33. Glosario de términos lingüísticos usados en el Diccionario Panhispánico de Dudas.



SOFTWARE ENGINEERING GLOSSARY

Software configuration management domain

baseline: 1. A description of a system and its components (configuration items) at a particular period of time, and any approved updates to the baseline. 2. A work product that has been placed under formal configuration management. See also *work product*.

baseline document: A system/software document that defines a work product that has been placed under configuration management. Examples are system specifications, requirements specifications, and design specifications.

build: An operational version of a software product incorporating a specified subset of the capabilities that the final product will include. Sometimes synonymous with *version*.

configuration: 1. The arrangement of a system or network as defined by the nature, number, and chief

port change processing and implementation status, and verify compliance with specified requirements. [IEEE Std. 610.12-1990]

configuration management system: The discipline of identifying the components of a continually evolving system for the purpose of controlling changes to those components and maintaining integrity and traceability throughout the life cycle.

control point (project control point): An agreed-on point in time or times when specified project agreements or controls are applied to the software configuration items being developed. Examples are an approved baseline or release of a specified document or code. [IEEE Std. 828-1998]

document control: The application of configuration management to the control of documents.

engineering change: An alteration in the configuration of a hardware/software configuration item or items, delivered, to be delivered, or under development, after formal establishment of their configuration identification. See also *engineering change proposal*.

engineering change proposal (ECP): A proposed engi-

ramento ■ thayer@csus.edu

Figura 2.34. Glosario de "Ingeniería del Software" (Thayer, 2003)

El uso de los glosarios y diccionarios en RI es muy limitado. En el caso de los diccionarios la relación con la RI es indirecta: se utilizan como fuente de conocimiento léxico para construir otros tipos de vocabularios adecuados a la RI (clasificaciones y taxonomías, tesauros, ontologías y bases de datos léxicas) (Zernik, 1991). Los glosarios, sin embargo, se han utilizado directamente como herramienta para acceder a los materiales, tanto para ampliar los mecanismos de búsqueda (Cuesta et. al, 2000), como para organizar y explorar información, contenidos y recursos (Tabata y Mitsumori, 2002).

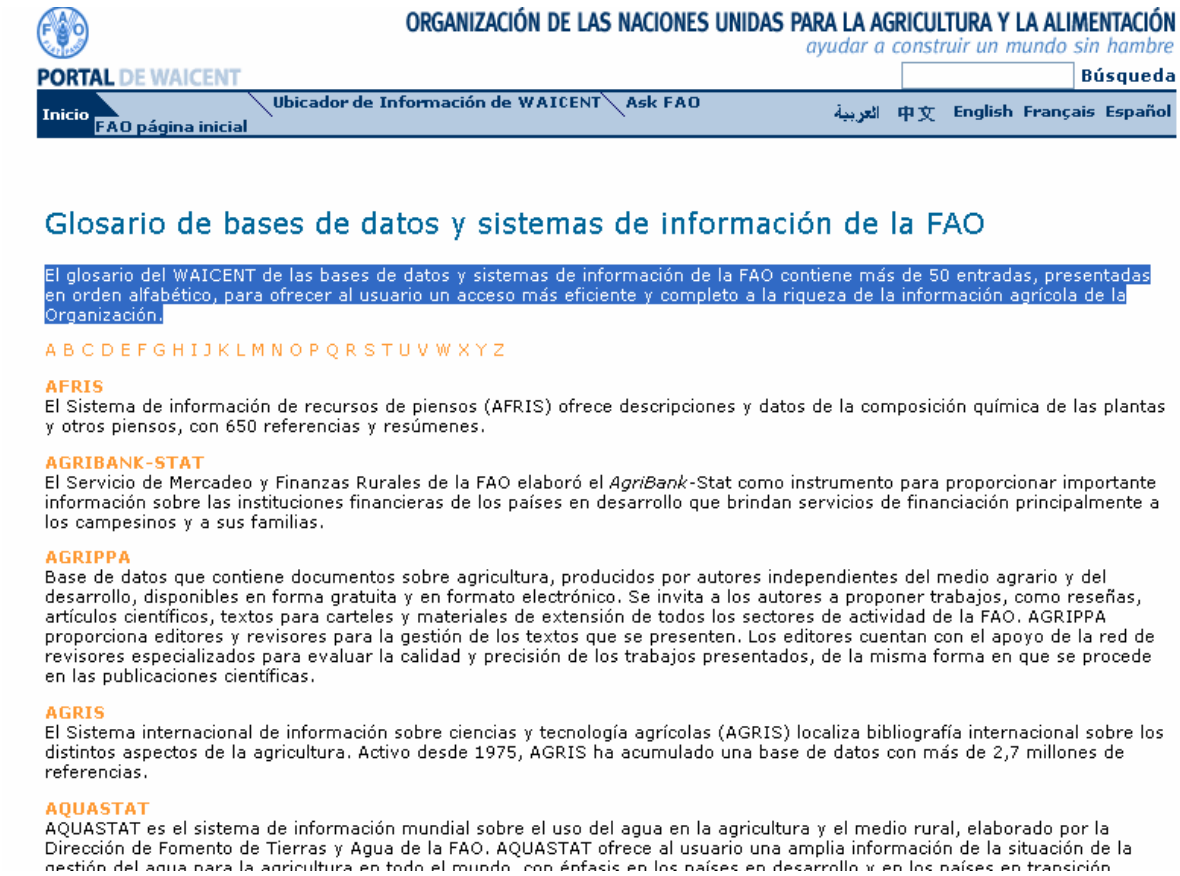


Figura 2.35. Glosario del portal WAICENT[82]

En este último caso destaca el uso de los glosarios como sistemas para describir el lenguaje de especialidad del contenido de las bases de datos documentales y sistemas de información (figura 2.35). El contenido de estas aplicaciones está, normalmente, descrito en alguna página Web de información o ayuda, que además incluye glosarios para definir el sentido preciso de la información, documentos o recursos almacenados (Hovy, et. al, 2003). Estos glosarios se construyen con algún esquema de datos

[82] El portal WAICENT es el centro de información agraria mundial de la FAO: http://www.fao.org/waicent/portal/glossary_es.asp

informático (por ejemplo texto HTML), (figura 2.36) por lo cual es posible acceder y procesar automáticamente la información que contienen.

```
<dt>
<a name="dt-argument"></a><b>Argument</b>
</dt>
<dd>A child of a presentation layout schema. That is, `A is an
argument of B' means `A is a child of B and B is a
presentation layout schema'. Thus, token elements have no arguments,
even if they have children (which can only be
<code>malignmark</code>).</dd>



<dt>
<a name="dt-attribute"></a><b>Attribute</b>
</dt>
<dd>A parameter used to specify some property of an SGML or XML element
type. It is defined in terms of an attribute name, attribute type, and a
default value.  A value may be specified for it on a start-tag for that
element type.</dd>
```

Figura 2.36. Detalle del código HTML de cada entrada del glosarios del MathML v. 2.0 del W3C[83]

Desde el punto de vista del usuario, los glosarios de los sistemas de información aportan no sólo un conjunto de términos del lenguaje natural para describir las entidades almacenadas, sino también la definición del sentido con el que se utilizan dichas palabras. De esta forma, el usuario puede comparar si lo que busca se corresponde con lo que se ha indexado y almacenado en el sistema, y, además, le ayuda a comprender y aprehender los conceptos del dominio de conocimiento que abarca la base de datos.

2.7. Resumen y conclusiones del capítulo

Para entender el papel de los vocabularios en la explotación de los recursos educativos digitalizados hemos revisado (i) su significado, que es diferente según los contextos disciplinares de aplicación; (ii) su naturaleza, que depende del tipo de vocabulario, (iii) la función en los sistemas de RI, y finalmente (iv) la función en la explotación de recursos didácticos digitalizados. De esta revisión podemos concluir que:

En primer lugar, en el contexto de la explotación de los recursos didácticos en entornos electrónicos vamos a definir el vocabulario como una herramienta de carácter lingüístico, informático y educativo que recoge y formaliza el conocimiento léxico de

[83] http://www.w3.org/TR/2001/REC-MathML2-20010221/appendixh.html

una lengua de especialidad para que pueda ser utilizado por las personas o los sistemas informáticos en la comprensión y enseñanza de los términos y nociones de una disciplina, y en la descripción y localización de recursos educativos digitalizados en la Web o en entornos e-learning[84] (Huynh et.al., 2005).

En segundo lugar, con respecto a la naturaleza de los vocabularios, consideramos que un vocabulario, en general, está formado por un conjunto de términos y sus descripciones. Cuando los términos se corresponden con un único significado se considera que el vocabulario es controlado. Las descripciones dependen del tipo de vocabulario, por ejemplo, (i) las listas no contienen descripciones; (ii) las taxonomías, tesauros y ontologías, sólo contienen relaciones semánticas entre los términos, que son de categorización y de hiponimia-hiperonimia en las clasificaciones y taxonomías, de equivalencia y asociativas en los tesauros y de tipología variada en las ontologías; (iii) los glosarios y diccionarios contienen descripciones expresadas en lenguaje natural orientadas al uso humano. Desde el punto de vista de explotación de recursos educativos en entornos electrónicos, los sistemas más simples para acceder al conocimiento por medio de la palabra son las clasificaciones y taxonomías y los tesauros.

En tercer lugar, respecto a la función de los vocabularios en los sistemas RI, éstos proporcionan un lenguaje preciso para describir y guiar la búsqueda de la información que se consulta. Se utilizan en la indexación y en la búsqueda de documentos. La eficiencia de los sistemas de RI se mide con los parámetros de exhaustividad y la precisión, que también se utilizan para medir la eficiencia de los vocabularios de RI, comparando los valores obtenidos por sistema RI utilizando y sin utilizar los vocabularios. Los vocabularios son eficaces en RI si son capaces de expresar el contenido de los documentos y las consultas de los usuarios de la misma forma. Por eso, en ciertas situaciones como por ejemplo, en las universidades, donde los profesores, investigadores y estudiantes son los autores, indexadores y usuarios del conocimiento, los vocabularios de referencia en un dominio de conocimiento que son considerados los más recomendables para la RI en ese dominio, no son más eficientes que otros vocabularios particulares más cercanos al lenguaje de especialidad compartido por los creadores y usuarios del dominio de conocimiento.

[84] En la web, los buscadores que incorporan los vocabularios como mecanismos de búsqueda son todavía escasos. Un ejemplo es el "Semantic Bank" en http://simile.mit.edu/bank/. Sin embargo, los repositorios de objetos de aprendizaje y los CMS ya incorporan vocabularios para la explotación de los contenidos que almacenan.

Finalmente, en cuarto lugar, respecto a la función de los vocabularios en la explotación didáctica de recursos digitalizados, éstos proporcionan un lenguaje preciso para indexar, buscar y representar las colecciones de recursos. Es recomendable utilizarlos con los metadatos, porque se combinan las ventajas de éstos, que son más sencillos de procesar, con las ventajas de los vocabularios, que proporcionan un lenguaje controlado y más expresivo para los valores de los metadatos. Sin embargo, actualmente, la aplicación de los metadatos y los vocabularios para la descripción y explotación de colecciones de recursos educativos digitalizados son escasos. Se apuntan como razones más importantes que, al igual que en RI, los vocabularios recomendados no se ajustan (i) al lenguaje de especialidad de los usuarios; (ii) a los contenidos de los recursos didácticos que deben describir e indexar; y (iii) al propósito de uso que, en la universidad es, a la vez, docente e investigador.

En definitiva, para mejorar eficacia de los vocabularios de explotación de recursos didácticos digitalizados en entornos académicos, éstos deben resolver los siguientes cometidos: (1) ajustarse a los recursos e información de la colección; (2) ajustarse al lenguaje de especialidad de los profesores que son autores y usuarios de los recursos; (3) ajustarse a los propósitos didácticos e investigadores; (4) servir de marco conceptual de referencia del conocimiento y trabajo de cada comunidad docente-discente específica; y (5) servir para aprender y comprender el lenguaje de especialidad y el conocimiento que expresan.

Capítulo 3

Los entornos virtuales de enseñanza y aprendizaje[1]

En 1962, R. Buckminster Fuller publica su visión de la enseñanza y el aprendizaje con el título Educación Automática, conjeturando que el futuro de la educación estará fuertemente condicionado por la tecnología, y se caracterizará por no tener límites geográficos o temporales:

> Get the most comprehensive generalized computer setup with network connections to process the documentaries that your faculty and graduate-student teams will manufacture objectively from the subjective gleanings of your vast new world- and universe-ranging student probes.

Las plataformas *e-learning*, plataformas educativas o entornos virtuales de enseñanza y aprendizaje (VLE[2]), constituyen, actualmente, esta realidad tecnológica creada en Internet y que da soporte a la enseñanza y el aprendizaje universitarios. En estos momentos podemos afirmar que su uso ha transformando una gran parte de los espacios de enseñanza tradicionales en espacios virtuales de enseñanza y aprendizaje[3] (EVA). Comprender, sin embargo, estas nuevas herramientas y saber cómo utilizarlas para mejorar la enseñanza y el aprendizaje es una tarea realmente compleja: un lenguaje confuso en el discurso del *e-learning* -con una gran cantidad de términos polisémicos y ambiguos-, y la contradicción entre la potencialidad teóricamente predicha para el *e-learning* a principios del años 2000 y los pobres resultados obtenidos, especialmente en términos económicos, en los siguientes años, convierten el *e-learning* en una cuestión aparentemente difícil de aplicar y poco rentable (Guri-Rosenblit 2005; Dondi 2008). Este capítulo aborda qué son, cómo son, cómo funcionan y qué aportan las plataformas *e-learning*. El propósito es contribuir a tener una visión más clara de los conceptos que consideramos claves para entender estas plataformas educativas y su uso. En este sentido, uno de los usos más frecuentes es la difusión y creación del conocimiento a través de los campus virtuales universitarios.

[1] Una versión previa de este capítulo ha sido publicada en el capítulo 2, páginas 45 a 75, con el título *Las plataformas e-learning para la enseñanza y el aprendizaje universitarios en Internet*, del libro "Las plataformas de aprendizaje. Del mito a la realidad", López Alonso, C. y Matesanz del Barrio, M. (eds.), editorial Biblioteca Nueva. Madrid 2009.

[2] Utilizaremos las siglas procedentes del inglés por su uso extendido en la bibliografía. VLE: Virtual Learning Environment.

[3] También llamados asignaturas virtuales, clases virtuales, aulas virtuales.

En ellos es donde surgen y se utilizan los tesauros y otros vocabularios como sistemas de referencia para la explotación -organización, descripción, consulta y utilización- del conocimiento y de los recursos didácticos.

3.1. Las plataformas e-learning y los espacios de aprendizaje

Una plataforma *e-learning*, plataforma educativa Web o entorno virtual de enseñanza y aprendizaje es una aplicación Web que integra un conjunto de herramientas para la enseñanza-aprendizaje en línea, permitiendo una enseñanza no presencial (*e-learning*) y/o una enseñanza mixta (*b-learning*), donde se combina la enseñanza en Internet con experiencias en la clase presencial (PLS Ramboll, 2004; Jenkins, Browne y Walker, 2005).

El objetivo primordial de una plataforma *e-learning* es permitir la creación y gestión de los espacios de enseñanza y aprendizaje en Internet, donde los profesores y los alumnos puedan interaccionar durante su proceso de formación. Un espacio de enseñanza y aprendizaje (EA) es el lugar donde se realiza el conjunto de procesos de enseñanza y aprendizaje dirigidos a la adquisición de una o varias competencias (Griffiths et al. 2004; López Alonso, Fernández-Pampillón y de Miguel, 2008). Los espacios de aprendizaje pueden ser[4] (i) las aulas de un centro educativo, en la enseñanza presencial; (ii) los sitios en Internet, en la enseñanza no presencial, virtual o *e-learning*; o (iii) la combinación de ambos, en la enseñanza mixta o *b-learning* (Britain y Liber, 2004).

Un *proceso de aprendizaje* se puede organizar mediante un *diseño de aprendizaje*[5]. En este caso, el diseño de aprendizaje (LD[6]) define y planifica la actuación de todos los elementos que participan en las relaciones didácticas: rol de profesores y alumnos, actividades que hay que realizar, escenarios[7], y relaciones entre roles, actividades y escenarios. Se puede comparar el espacio de aprendizaje con un teatro[8] (edificio o sitio para el ocio) en el que se representan obras, que son los procesos de aprendizaje,

[4] Algunos autores dan un sentido habitual de "espacio cognitivo" al concepto de EA, pero entendiendo que incluye el espacio físico donde se realiza el proceso educativo (Banyard y Underwood, 2008).

5 Es el caso cuando la enseñanza se realiza únicamente en EA virtuales, e-learning; pero es posible encontrar EA virtuales sin un diseño de aprendizaje, bien porque se utilizan esporádicamente como complemento de la enseñanza presencial o bien porque se conciben como EA "libres" donde los alumnos pueden explorar o aprender según sus propios esquemas de organización del aprendizaje.

6 De nuevo utilizaremos la sigla del inglés, Learning Desing, por estar ampliamente difundida.

7 El escenario es una parte del espacio de aprendizaje donde se realiza un único proceso de aprendizaje.

8 Este símil está inspirado en Rob Kopper para explicar el concepto de unidad de aprendizaje, pero con un significado más amplio que el utilizado en la metodología que propone (Koper, 2005).

con un guión, que es el diseño de aprendizaje. Finalmente, el escenario es la zona del teatro donde se representan la obra[9]. La figura 3.1 muestra gráficamente la relación entre entornos virtuales de aprendizaje, espacios virtuales de aprendizaje, escenarios y diseños de aprendizaje.

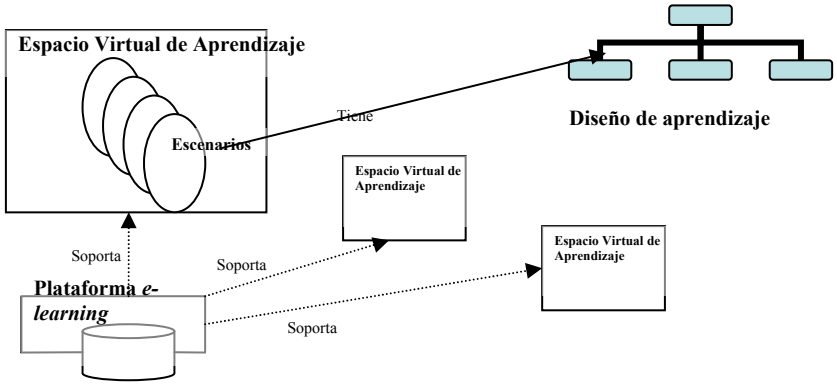


Figura 3.1. Modelo conceptual de espacio virtual de aprendizaje

En el ámbito no académico, las administraciones, empresas, compañías y otras organizaciones utilizan las plataformas *e-learning* para la formación, entrenamiento o perfeccionamiento permanente de sus empleados, con un enfoque instruccional. El fin es ofrecer a su personal una herramienta de perfeccionamiento profesional permanentemente accesible y de bajo coste. A pesar de que éste ha sido el enfoque original de las plataformas, en el ámbito académico, y específicamente en el contexto universitario, el objetivo de uso cambia hacia la búsqueda y aplicación de modelos y métodos educativos más eficaces para profesores y alumnos. Actualmente, el uso de las plataformas en las universidades está muy generalizado y su explotación se realiza desde múltiples aproximaciones pedagógicas, especialmente en aquellos centros con un modelo de aplicación centrado en el profesor e, incluso, de formas no previstas en la concepción original de estas plataformas[10] (Dondi, 2008; Fernández-Valmayor et. al. (eds), 2008). Esta explotación está produciendo un avance no sólo en las propias plataformas, a las que se les demandan más funciones, más flexibilidad y mayor robustez, sino también en la propia actividad docente universitaria, que está experimentando un proceso de innovación tecnológica y metodológica.

[9] También es posible considerar la reutilización de un mismo escenario para llevar a cabo procesos de aprendizaje diferentes.
[10] Extendiéndose su uso más allá de la docencia, en actividades de investigación y de gestión académica.

En la próxima sección, revisaremos el uso de las plataformas *e-learning* desde un modelo de explotación ampliamente utilizado en las universidades: los campus virtuales.

Respecto a la funcionalidad de las plataformas educativas distinguimos entre las que son de carácter general y las específicas. En el primer caso, una plataforma se considera de carácter general cuando es "pedagógicamente neutra"[11] y no está orientada hacia el aprendizaje de una materia concreta o hacia la adquisición de una competencia en particular o a la realización de una función específica. En este caso, los sistemas software más utilizados son los sistemas de gestión del aprendizaje (Learning Management Systems) o LMS[12]. Como ejemplos de LMS de código abierto podemos mencionar Moodle[13], .LRN[14] o el reciente Sakai[15] y, entre los sistemas comerciales, el más extendido es Blackboard-WebCT[16], e-College[17] o Desire2Learn[18].

Los LMS permiten crear y gestionar múltiples espacios virtuales de aprendizaje, privados para cada grupo de estudiantes y profesores. Estos EA se crean, normalmente, incorporando a una plantilla que puede personalizarse un conjunto de herramientas que el diseñador, el profesor o el administrador del sistema consideran necesarias para llevar a cabo los procesos de aprendizaje (figura 3.2).

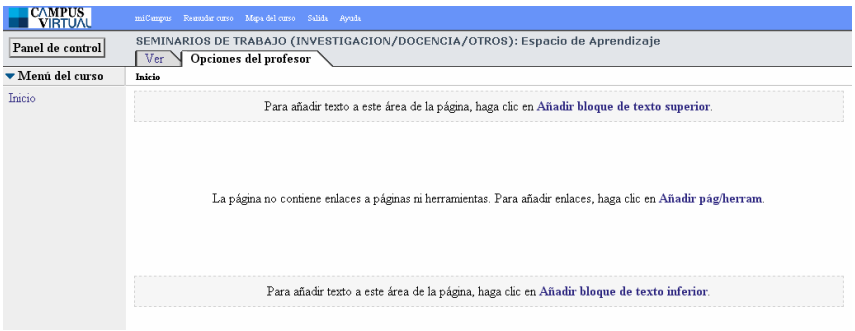


Figura 3.2. Plantilla vacía de los espacios de aprendizaje del LMS WebCT 4.0[19]

11 Esta caracterización está muy cuestionada porque, desde el punto de vista de rentabilidad académica, no parece que se pueda considerar "neutra"(Conole y Fill, 2005); lo cierto es que permite utilizar múltiples metodologías didácticas.

12 En este caso, utilizaremos las siglas en inglés, LMS, porque son habitualmente utilizadas en la bibliografía en inglés y en español. Conviene también tener en cuenta que ciertos autores los denominan Course Management Systems (CMS), pero estas siglas pueden confundirse con los Content Management Systems, que son plataformas específicas para la gestión de contenidos.

13 http://moodle.org/

14 http://dotlrn.org/

15 http://sakaiproject.org/

16 http://www.blackboard.com/

17 http://www.ecollege.com/

18 http://www.desire2learn.com/

19 http://www.webct.com/ce4

El conjunto de herramientas de un LMS (figura 3.3) permite realizar cinco funciones principales: (i) la administración del EA; (ii) la comunicación de los participantes; (iii) la gestión de contenidos; (iv) la gestión del trabajo en grupos; y (v) la evaluación. Aunque cada LMS tiene su propio conjunto de herramientas[20] destacamos, a continuación, algunas de las más comunes para tener una visión general de cómo se puede implementar cada una de estas cuatro funciones.

(i) *Administración.* Estas herramientas deben facilitar, en primer lugar, las operaciones de gestión de usuarios: altas, modificaciones, borrado, gestión de la lista de clase, la definición de roles y el control y seguimiento del acceso de los usuarios al EA o a sus diferentes partes. En segundo lugar, la gestión de los EA: creación, modificación, visibilidad y eliminación del EA o de sus partes como, por ejemplo, configuración del formato de la plantilla, incorporación, eliminación o definición de criterios de visibilidad de las herramientas.

(ii) *Comunicación.* Las herramientas de comunicación permiten la interacción entre profesores y alumnos. Puede ser asíncrona[21] con el correo electrónico, los foros, el calendario y los avisos; o síncrona, con las charlas (chats) o la pizarra electrónica. Estas herramientas permiten todos los sentidos de interacción: del profesor hacia alumnos, de los alumnos hacia profesor, de alumno con alumnos, alumnos entre sí, o todos con todos.

(iii) *Gestión de contenidos*. Para la gestión de contenidos los LMS disponen de un sistema de almacenamiento y gestión de archivos que permite realizar operaciones básicas sobre ellos, como visualizarlos, organizarlos en carpetas (directorios) y subcarpetas, copiar, pegar, eliminar, comprimir, descargar o cargar archivos en el EA. Además, suelen incorporar algún sistema para la publicación organizada y selectiva de los contenidos de dichos archivos, y alguna herramienta muy básica para la creación de contenidos[22]. No tienen restricciones respecto a los tipos de archivos, pero para su visualización es necesario que el usuario tenga instalada localmente en el ordenador desde el que hace la consulta, la aplicación apropiada[23].

---

[20] Puede consultarse en EduTools una evaluación comparativa de LMS (que denomina Course Management Systems): http://www.edutools.info/

[21] En este tipo de comunicación los mensajes quedan almacenados y están disponibles para todos los participantes sin límite de tiempo, mientras que en la comunicación síncrona los mensajes se producen y reciben en un determinado momento y están disponible mientras dure la interacción.

[22] Básicamente, editores de textos o de texto html.

[23] Por ejemplo, para la visualización de un archivo ppt el ordenador de consulta debe tener instalado el MS Power Point.

(iv) *Gestión de grupos*. Estas herramientas permiten realizar las operaciones de alta, modificación o borrado de grupos de alumnos y la creación de "escenarios virtuales" para el trabajo cooperativo de los miembros de un grupo. Estos escenarios de grupo incluyen directorios o carpetas para el intercambio de archivos, herramientas para la publicación de los contenidos, y foros o chats privados para los miembros de cada grupo.

(v) *Evaluación*. Las herramientas para la evaluación permiten la creación, edición y realización de ciertos tipos de tests[24], anónimos o nominales, trabajos, la autocorrección o la corrección (con realimentación), la calificación y publicación de calificaciones, y la visualización de información estadística sobre los resultados, así como el progreso de cada alumno.

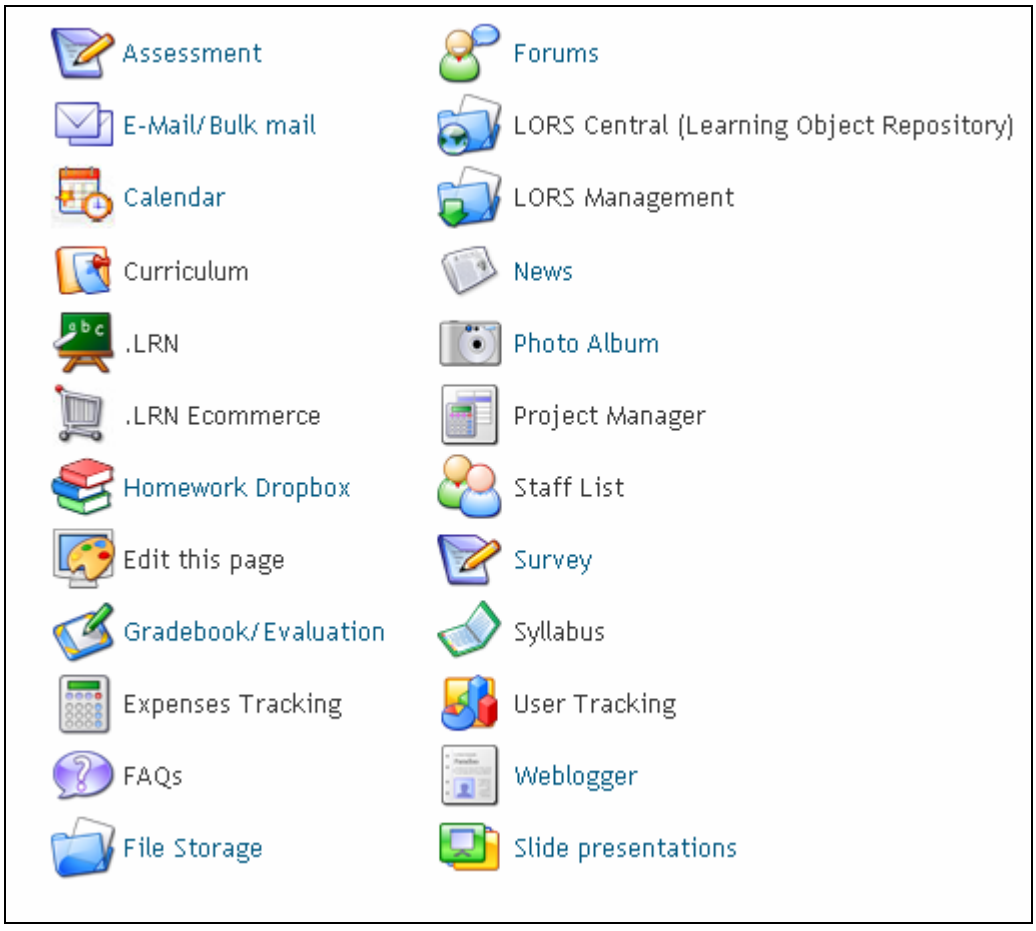


Figura 3.3. Conjunto de herramientas del LMS de código abierto .LNR

[24] El tipo de pregunta (opción simple, múltiple, respuesta corta, larga, calculada, relacional, etc.) cambia de un LMS a otro.

Frente a las plataformas educativas genéricas se encuentran las plataformas específicas, con el objetivo de mejorar la eficacia y eficiencia académica -mejor y más rápida enseñanza y aprendizaje-, especializándose en determinadas áreas de conocimiento o completando la funcionalidad de las plataformas genéricas. Así, encontramos plataformas especializadas en (i) un dominio (competencia o materia) concreto; (ii) un modelo y/o metodología de aprendizaje específico, o finalmente, (iii) una tarea específica. Estas plataformas construyen y gestionan los EA siguiendo unos criterios específicos del dominio. En la mayoría de los casos, la propia interfaz de la plataforma es el único EA posible (figura 3.4).

Un ejemplo paradigmático del primer caso, las plataformas específicas para el desarrollo de una destreza o el aprendizaje de una materia concreta, son las plataformas orientadas al aprendizaje de las lenguas (figura 3.4). Estos sistemas integran las herramientas que se adaptan a las metodologías específicas de enseñanza de esa competencia. Los EA suelen estar ya definidos, aunque se permite la personalización de la plantilla y la elección de la lengua de interacción. Las herramientas utilizadas habitualmente son las de (i) comunicación síncrona multimedia (por ejemplo, videoconferencia); (ii) almacenamiento masivo y clasificación de recursos didácticos digitalizados, (por ejemplo, repositorios de archivos de vídeo, sonido, hipertextos y textos); (iii) construcción de vocabularios (por ejemplo, diccionarios y tesauros); (iv) materiales educativos multimedia e interactivos (por ejemplo, gramáticas, ejercicios de audio, video y texto), (v) trabajo colaborativo (por ejemplo, blogs, wikis, podcasting[25]); (vi) soporte multilingüe (por ejemplo, interfaz en múltiples lenguas); y (vii) definición de los perfiles de los participantes, de votación, y de publicación de trabajos de alumnos (López Alonso y Séré 2005; Monti, San Vicente y Preti, 2006).

[25] Las herramientas de podcasting permiten la creación de archivos de sonido (generalmente en formato mp3 o AAC, y en algunos casos ogg) y de vídeo (llamados videocasts o vodcasts) y distribuirlos mediante un archivo RSS que permite suscribirse y usar un programa que lo descarga de internet para que el usuario lo escuche en el momento que quiera, generalmente, en un reproductor portátil (Wikipedia, 2008: http://es.wikipedia.org/wiki/Podcasting).

Figura 3.4. GALANET, plataforma de formación en la intercomprensión entre lenguas románicas[26]

Estas plataformas se diseñan con sólidos fundamentos didácticos que han sido previamente experimentados en entornos reales y que han demostrado mejorar la eficacia de los procesos de aprendizaje. Su desarrollo implica equipos multidisciplinares de informáticos y expertos en el dominio y en la enseñanza del dominio.

En el segundo caso, plataformas orientadas a un modelo o método de aprendizaje específico. Uno de los ejemplos más emblemáticos, y de muy reciente aparición, son los entornos personales de aprendizaje (Personal Learning Environments o PLE). Estas plataformas no han sido concebidas estrictamente como plataformas educativas (y realmente no existen como tales), pero están basadas en el modelo de aprendizaje socioconstructivista en el que el aprendiente es protagonista de su propio aprendizaje, cooperando y colaborando con el grupo para construir nuevos conocimientos. Surgen como un fenómeno más de la próxima versión de la web, la web semántica o web 2.0, en la que los usuarios son creadores, además de consumidores de información (Schaffert y Hilzensauer, 2008). Netvibes[27] podría ser, actualmente, el mejor ejemplo. Los PLE están formados por una plantilla, que puede personalizarse, y un conjunto de 'herramientas de software social[28]' que permiten a los participantes: (i) la creación de

[26] http://www.galanet.eu/
[27] http://www.netvibes.com/
[28] El software social puede definirse como el software que conecta a las personas y asegurar su colaboración y comunicación.

su propio EA; (ii) la creación y publicación colaborativa de contenidos, por ejemplo, wikis, weblogs, podcasting; (iii) la integración, el almacenamiento, clasificación e indexado de múltiples fuentes de información y datos, por ejemplo, del.icio.us3, Flickr4, YouTube5; (iv) la comunicación multimedia e instantánea, por ejemplo, Skype, AIM1, ICQ2; y (v) la creación de sus propias comunidades (MySpace6, Facebook7, LinkedIn8), por ejemplo, Netvibes.

En el tercer caso, las plataformas *e-learning* con funciones más específicas, se incluyen sistemas, como los sistemas de gestión de contenidos -Content Management Systems- (CMSs), los sistemas de gestión del aprendizaje y contenidos -Learning Content Management System- (LCMS)[29], los sistemas de gestión de secuencias de actividades -Learning Activities Management Systems-, y los sistemas síncronos de gestión del aprendizaje, de muy reciente aparición. El objetivo de estos sistemas es completar las capacidades de los LMS, bien integrándose con el LMS, bien creando EA específicos, pero accesibles mediante un hiperenlace desde los EA principales de los LMS o, simplemente, creando EA sólo con las funcionalidades específicas de la plataforma.

Los CMS son aplicaciones que permiten la creación, almacenamiento indexado, clasificación, publicación y gestión multiusuario y concurrente del ciclo de vida de los contenidos. Complementan las capacidades de los LMS, limitadas al mero almacenamiento en directorios y a la publicación. Su inclusión como plataformas *e-learning* es discutible porque su funcionalidad está limitada a la creación y gestión de espacios de contenidos. Sin embargo, en la medida en que los contenidos y la creación de contenidos es un recurso y una actividad principal de la enseñanza -los actuales LMS carecen de esta capacidad-, y teniendo en cuenta que incluyen otras herramientas, como la creación y gestión de espacios personales, comunicación (foros, e-mail) o calendario, pueden considerarse plataformas *e-learning* específicas de contenidos. Phone[30], Silva[31] o Drupal[32] son algunos buenos ejemplos. Estas plataformas junto a los LCMS incorporan vocabularios más o menos complejos (listas

[29] Ciertamente, existe bastante confusión en el uso de estos acrónimos que son polisémicos (la primera C puede significar Content o Course). La interpretación que utilizamos en este capítulo es Content y está en consonancia con la utilizada por el W3C (www.w3c.org) y el IMS Global Learning Consortium (www.imsglobal.org).
[30] http://phone.org
[31] http://www.infrae.com/products/silva
[32] http://www.drupal.org/

de categorías, taxonomías o tesauros) para el indexado, clasificación y navegación de sus contenidos.

Los LCMS, al igual que los CMS, proporcionan una gestión de contenidos, pero orientada al *e-learning* e integrando, generalmente, los estándares de producción de contenidos educativos reutilizables IMS (IMS, 2002) y SCORM (SCORM, 2004). Estos sistemas pueden estar integrados en un LMS proporcionando, además de un sistema de autoría, un repositorio de objetos de aprendizaje que el profesor puede utilizar y reutilizar para sus cursos en el LMS (Hall, 2007). Un ejemplo es el sistema ATutor[33].

Los sistemas de gestión de secuencias de actividades tienen como objetivo la construcción de EA instruccionales. Incluyen, además de algunas herramientas básicas de un LMS, herramientas para la definición, creación y actualización de secuencias de actividades de aprendizaje, así como el control, seguimiento y la evaluación. LAMS[34] es probablemente la única plataforma *e-learning* de estas características utilizada en la enseñanza-aprendizaje real, que puede ser, además, integrada en otros LMS[35]. *CopperCore*[36] es una aplicación software para diseñar y ejecutar secuencias de aprendizaje definidas con el estándar IMS-LD. Esta aplicación es de código abierto y está preparada para integrarse en plataformas *e-learning*.

Las plataformas *e-learning* síncronas crean EA donde profesores y alumnos interaccionan en tiempo real, viéndose y escuchándose como si de una clase presencial se tratase[37] (figura 3.5). El espacio de aprendizaje contiene herramientas (i) para la lectura, escritura participativa, como, por ejemplo, una pizarra electrónica; (ii) para la comunicación síncrona por audio, vídeo y chat; administración del EA (altas y bajas de usuarios, control y seguimiento de su actividad) y (iii) intercambio de archivos, block de notas personal del estudiante, gestión de grupos, chats y evaluación.

Además de las plataformas educativas, existe un amplio abanico de herramientas satélites que complementan la funcionalidad de las plataformas *e-learning*. Son aplicaciones informáticas independientes, que no tienen como fin la creación y gestión de EA, pero que forman parte del software para el *e-learning*. Destacamos por

[33] http://www.atutor.ca/
[34] http://lamsfoundation.org/
[35] Puede consultarse información actualizada en: http://lamsfoundation.org/integration/
[36] http://coppercore.sourceforge.net
[37] Estas herramientas utilizan tecnología multidifusión (multitast) para la emisión de audio y vídeo en tiempo real.

su utilidad: (i) las herramientas de autoría de exámenes de corrección automática (tests) como HotPotatoes[38] o Respondus[39], que permiten crear múltiples tipos de preguntas y exámenes en formatos estándares y propietarios[40] para la importación a los LMS; (ii) las herramientas de autoría de contenidos, como courseGenie, eXe, Lectora; y (iii) los repositorios de recursos didácticos digitalizados.

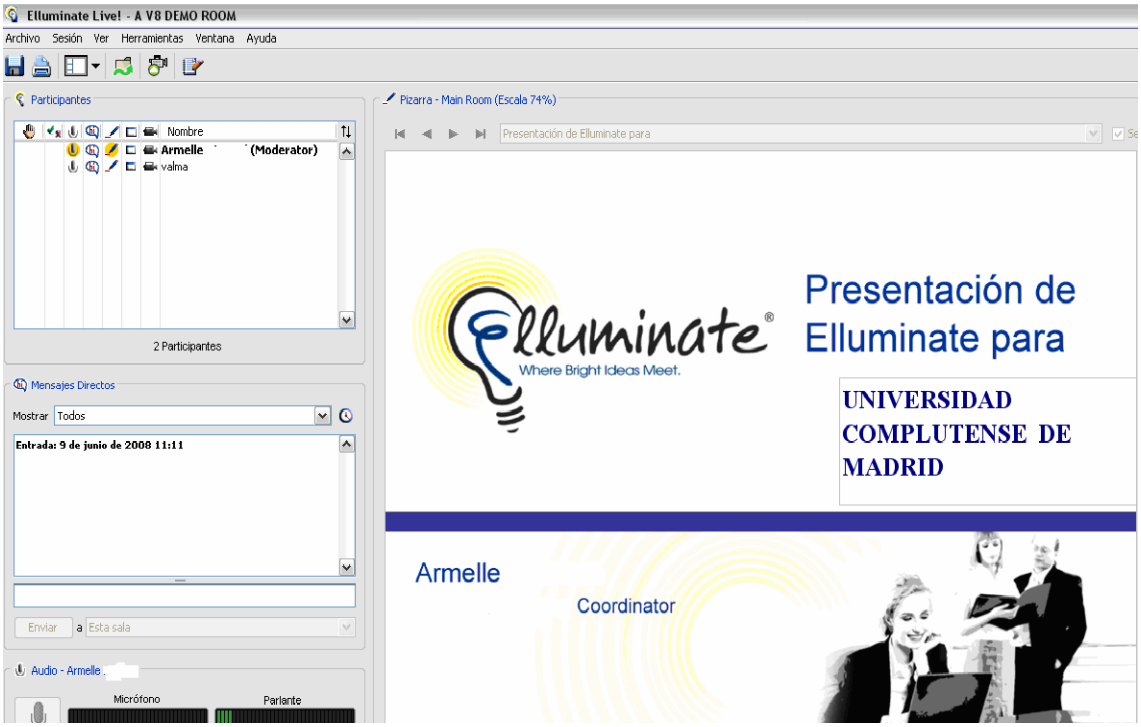


Figura 3.5. Espacio de aprendizaje de la plataforma e-learning síncrona Elluminate Live![41]

Un repositorio educativo es un contenedor en línea de recursos educativos digitalizados creados y compartidos por un grupo de usuarios (figura 3.6). En el *e-learning* existe actualmente un conjunto de normas técnicas estándares para representar y estructurar el recurso[42] y para documentarlo, de forma que puedan ser compartidos y reutilizados los recursos con mayor facilidad, rentabilizando su construcción (Hernández, 2003). Los repositorios de este tipo se llaman repositorios

[38] http://hotpot.uvic.ca/ y página con la información en español http://platea.pntic.mec.es/~iali/CN/Hot_Potatoes/intro.htm.
[39] http://www.respondus.com/
[40] El formato es *propietario* o cerrado cuando el archivo sólo puede ser generado, reconocido y utilizado por la aplicación software particular que tiene la patente o derechos de autor. Por ejemplo, los documentos con extensión ppt son propietarios en la medida que se generan y utilizan con el procesador MS Power Point. El opuesto es un formato abierto, que no impone restricciones de uso. Los formatos *estándares* son formatos abiertos y consensuados.
[41] http://www.elluminate.com/
[42] Con los estándares IMS Content Package (http://www.imsglobal.org/content/packaging/) y SCORM (http://www.adlnet.gov/scorm/index.aspx)

de objetos de aprendizaje[43] (OA) y almacenan, además de los propios OA, una descripción de las características de cada uno. Esta descripción se conoce como metadatos, cuando está formada por un conjunto de pares de atributos y valores como, por ejemplo, autor: Pepe Jiménez; materia: matemáticas/conjuntos; o nivel educativo: secundaria. Estos metadatos permiten a los usuarios –registrados o no- realizar, de forma más precisa, la búsqueda, selección y recuperación de los objetos, comparando los valores de los atributos del OA con sus necesidades. La búsqueda puede ser simple, avanzada o realizarse navegando en un índice de materias o disciplinas basado en un vocabulario (figura 3.6). Pueden, además, incluir otras funcionalidades como, por ejemplo, la posibilidad de contribuir con nuevos OA; participar en la evaluación en línea de la calidad de los materiales almacenados; espacios personales para archivar OA o enlaces a los preferidos; perfiles de usuario basados en OA previamente seleccionados para dirigir posteriores búsquedas o avisar de nuevos OA ajustados al perfil; foros y chats de usuarios; y el soporte a comunidades de usuarios con perfiles similares (Neven y Duval, 2002). Como ejemplos destacamos el repositorio europeo ARIADNE[44] y el reciente repositorio español Agrega[45], que es un proyecto del Ministerio de Educación y Ciencia, Red.es y las Comunidades Autónomas[46] para promover la construcción, uso y compartición de recursos educativos digitalizados de calidad en España.

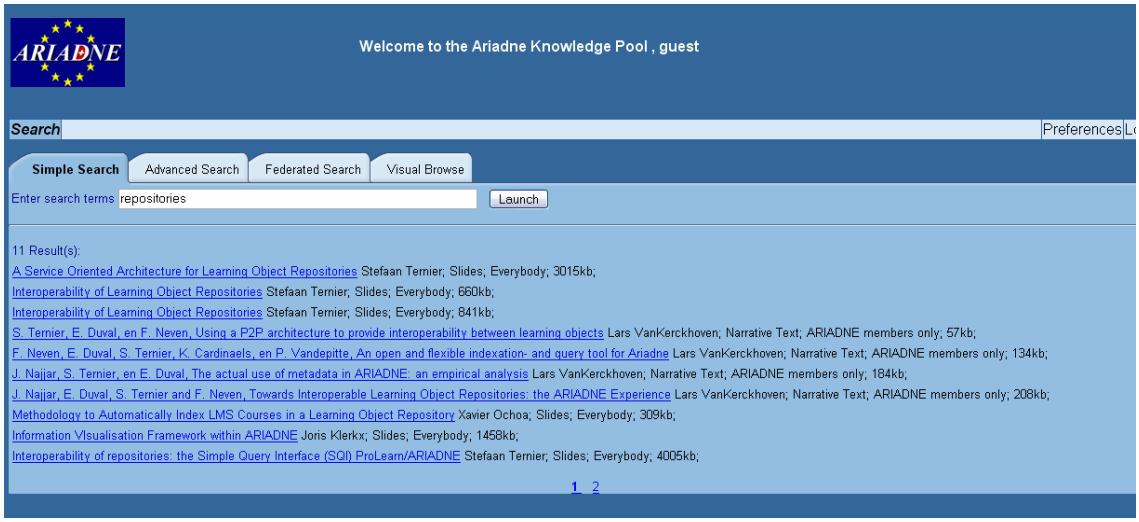


Figura 3.6. Repositorio Europeo de Objetos de Aprendizaje ARIADNE

[43] Learning Objects.
[44] www.ariadne-eu.org
[45] http://www.proyectoagrega.es
[46] http://redes.agrega.indra.es

Las plataformas *e-learning*, a modo de resumen, permiten la creación y uso de los espacios de aprendizaje en la Web, a los que se accede simplemente con un navegador.

Las plataformas de carácter general permiten crear múltiples espacios diferentes a partir de una plantilla y un conjunto de herramientas. El diseñador del EA selecciona y organiza estas herramientas conforme a la definición, implícita o explícita, del diseño del proceso de aprendizaje fundamentado, a su vez, en modelos y métodos didácticos. Los procesos de enseñanza y aprendizaje se realizan en estos EA con la participación de profesores y alumnos, y las plataformas se encargan de la ejecución, control y seguimiento de la actividad de cada participante.

Las plataformas específicas, a diferencia de las generales, tienen ya definidos los EA, aunque permiten cierta personalización, con una plantilla y un conjunto de herramientas seleccionadas conforme a un método didáctico fundamentado y experimentado o bien conforme a la funcionalidad más específica que proveen. Estas plataformas son más eficaces que las genéricas en sus dominios concretos de aplicación, sin embargo, presentan limitaciones que tienen que valorarse a la hora de decidir el tipo de plataforma educativa más adecuada. Estas limitaciones son los elevados costes de desarrollo y mantenimiento respecto del restringido ámbito de uso, baja rentabilidad, y la difícil reutilización de sus componentes debido a la fuerte dependencia del dominio.

Las herramientas satélite no forman parte de las plataformas y no crean ni gestionan EA, pero apoyan a los EA soportados por las plataformas *e-learning* en alguna de sus funciones. De estas herramientas destacamos los repositorios de OA, porque permiten la creación colaborativa, la compartición y la reutilización de los recursos educativos de calidad, que son muy caros de construir y mantener, pero que son muy eficaces para la enseñanza y aprendizaje en entornos virtuales.

La complejidad de este panorama tecnológico de plataformas y herramientas tiene, entre otras consecuencias, dos que son particularmente significativas. La primera, es la necesidad de encontrar soluciones más simples que abaraten los costes de implantación y mantenimiento. Actualmente, los esfuerzos se dirige en dos direcciones: a) la búsqueda de mecanismos de integración entre los actuales

sistemas[47], y b) el desarrollo de versiones o nuevas plataformas, cada vez más completas, que integran capacidades de las plataformas específicas o de las herramientas satélites[48]. La segunda consecuencia es la necesidad de soporte institucional para poder incorporar el uso de las plataformas *e-learning* a la enseñanza-aprendizaje. No es factible pensar que un profesor o equipo de profesores puede abordarlo aisladamente por: (i) el alto coste económico de implantación y, sobre todo, de mantenimiento de plataformas y herramientas; (ii) los recursos de personal -técnico y docente-, necesarios para asegurar el funcionamiento correcto de la infraestructura técnica e informática; y (iii) la dedicación que requeriría – además de una formación especializada en tecnologías e informática- el seguimiento de los rápidos y continuos avances en esta área, que evite la obsolescencia de los EA y sus componentes. La pregunta que surge entonces es ¿cómo pueden beneficiarse los profesores y alumnos del uso de las plataformas *e-learning* a un coste razonable para su institución y para ellos?

3.2. El uso de las plataformas *e-learning* en los campus virtuales universitarios

En la actualidad, el mecanismo más extendido y de mayor éxito para el uso de las plataformas *e-learning* en la enseñanza y el aprendizaje universitario es el campus virtual (CV). Un campus virtual puede definirse como el lugar para la enseñanza, aprendizaje e investigación creado mediante la confluencia de múltiples aplicaciones de la Tecnología la Información y las Comunicaciones (TIC): internet, la web, comunicación electrónica, video, video-conferencia, multimedia y publicación electrónica (Van-Dusen, 1997)[49]. Esta definición, ya clásica, puede actualizarse considerando la aparición, posteriormente, de las plataformas *e-learning*, que integran las herramientas TIC, a las que se refiere Van Dusen, en una única aplicación con fines educativos. En este momento podemos definir el CV como el espacio en internet creado con aplicaciones web, principalmente plataformas *e-learning*, con un propósito educativo. Otro término muy relacionado con CV es el de universidad

[47] Definiendo estándares para la interoperatividad de herramientas como, por ejemplo, IMS Tools Interoperability Guidelines Compativility, o mediante acuerdos entre las compañías u organizaciones de productos *e-learning* (por ejemplo, LAMS y Moodle; LAMS y Blackboard).

[48] Por ejemplo, Blackboard y Moodle integran, en sus últimas versiones, funciones de los repositorios de OA.

[49] Estos espacios también pueden llamarse aula virtual, universidad en línea (on-line), eCampus, eUniversidad, ciberaula y campus tecnológico.

virtual. Tiene un significado más amplio, ya que se refiere no sólo a los espacios para la enseñanza, aprendizaje e investigación, sino también a los espacios para la administración y organización de todas las actividades y procesos de una universidad (PLS Ramboll, 2004), (Epper y Gran, 2004). Los sistemas software de soporte son también más generales y se denominan Managed Learning Environment (MLE)[50] (IMS, 2007). Nosotros utilizaremos para este último caso el término universidad virtual y reservamos el de campus virtual para la instancia educativa.

En cualquiera de los casos, las plataformas *e-learning* constituyen el soporte técnico de los CV concebidos bien como el conjunto de espacios de enseñanza y aprendizaje de una institución, la Universidad, o bien como un subconjunto del total de espacios virtuales de esa institución, que están dedicados exclusivamente a la enseñanza, el aprendizaje y la investigación. En el primer caso, las plataformas *e-learning* se utilizan como sistemas autónomos, mientras que en el segundo caso están integradas en los MLE (Epper y Garn, 2004).

El propósito de un CV es que los profesores y alumnos puedan aprovechar las funciones que les ofrecen para optimizar su trabajo docente y discente. Las posibilidades de uso no son siempre las mismas, dependen de cómo sea el CV en su universidad.

3.2.1. La forma de los campus virtuales

La forma de un CV viene determinada por el modelo conceptual y físico que se adopte. Este modelo es definido a nivel institucional y depende de varios parámetros, como son: (i) las finalidades institucionales, (ii) la infraestructura organizativa y de funcionamiento, (iii) los recursos, y (iv) el contexto político, social y económico de la institución (Romiszowski, 2004; PLS Ramboll, 2004; Epper y Garn, 2004). Para facilitar el análisis de los modelos de CV a nivel conceptual, consideramos estos parámetros agrupados en tres dimensiones[51]: 1) institucional, 2) tecnológica y 3) didáctica (figura 3.7).

[50] Los MLE incluyen a las plataformas *e-learning* o VLE.

[51] El número de dimensiones propuestas para el análisis de los proyectos e-learning varía según los autores y el propósito del análisis. Por ejemplo, Khan (2005) para la evaluación de los proyectos propone siete dimensiones, fundamentalmente, de tipo tecnológico-informático; para tomar decisiones acerca del modelo e-learning McGraw (2001) toma cuatro y Epper, y Garn (2004) tres. Todas las propuestas incluyen las tres dimensiones básicas: institucional, tecnológica y didáctica, que son las definidas para los análisis de modelos a nivel conceptual

1) La dimensión institucional. Se refiere a la definición del propósito que tiene el CV, la estructura organizativa, las normas de funcionamiento, la estrategia de difusión y el apoyo o soporte previsto para desarrollar las dimensiones tecnológicas y didácticas.

2) La dimensión tecnológica. Se refiere a la definición de la infraestructura informática, tecnológica y de comunicaciones, y a los recursos económicos y de personal necesarios para su desarrollo y mantenimiento.

3) La dimensión didáctica. Se refiere a la definición del tipo de participación de los usuarios –profesores, alumnos, diseñadores-, a las metodologías didácticas que se van a promover, la formación y el soporte, así como las políticas de promoción de uso del *e-learning*.

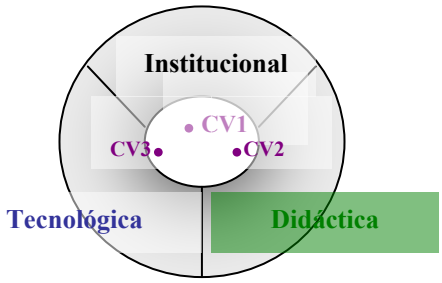


Figura 3.7. Dimensiones para el análisis de los modelo de CV

Los valores escogidos en cada una de estas dimensiones[52] son los que definen un modelo de campus virtual. En nuestro análisis encontramos cuatro tipos de modelos, en función de la dimensión que priorizan según estén centrados en: (i) la tecnología, (ii) la institución, (iii) el alumno y (iv) el profesor.

3.2.1.1. Modelo centrado en la tecnología

En este modelo se da prioridad a la plataforma *e-learning* y la infraestructura informática y de comunicaciones frente a las otras dimensiones[53]. Esta es la estrategia utilizada para la construcción de los primeros CV y que todavía sigue vigente incluso como criterio principal para el análisis de los modelos de CV (PLS Ramboll, 2004). Los CV centrados en la tecnología invierten importantes recursos económicos y humanos en la creación, evaluación, mantenimiento y actualización de sus propias

[52] Las dimensiones son, a su vez, multiparamétricas (véase la nota 50), por lo que se representan como un sector de círculo que es un continuo de puntos.
[53] En la figura se representa con la etiqueta CV3, más cercano a la dimensión tecnológica que los otros modelos.

plataformas *e-learning*. El papel del profesor en la construcción y evolución del CV es mínimo, y se reduce a impartir docencia en las plataformas desarrolladas por la institución. El grado de participación de los profesores depende de la política de compensaciones de la universidad, que normalmente es escasa o nula. El grado de participación del alumno depende de la participación del profesor, aunque siempre utiliza las herramientas de comunicación, aún sin la participación docente, para ayudarse durante su aprendizaje.

Este modelo está siendo muy cuestionado por sus resultados poco satisfactorios. Los estudios indican que son poco rentables en términos económicos, dados los altos costes respecto del número de usuarios y, en términos didácticos, tienen altos costes respecto del número de alumnos que completan con éxito los cursos y respecto del grado de satisfacción de los profesores (Romiszowski, 2004; Guri-Rosenblit, 2005). La causa de este fracaso está en la pregunta formulada por Romiszowki: "si la tecnología es la solución, ¿cuál es el problema?". Este modelo prioriza la dimensión tecnológica, pero sin una reflexión previa sobre cuáles son los propósitos y funciones que debe cumplir esta tecnología desde el punto de vista institucional y didáctico.

3.2.1.2 Modelos centrados en la institución

Estos modelos dan prioridad a los objetivos de la institución, normalmente económicos o político-sociales, frente a planteamientos didácticos o a necesidades tecnológicas[54] (Guri-Rosenblit, 2005). Este modelo es adoptado frecuentemente cuando la institución es una universidad virtual (D'Antony, 2003). Se trata de universidades "abiertas"[55], con una enseñanza a distancia. Es el caso, entre otras, de la Universidad Abierta de Cataluña (UOC)[56], la Universidad Nacional de Educación a Distancia (UNED)[57], la Universidad Abierta de Holanda (Open University Netherlands)[58], la Universidad de Athabaska en Canadá[59] (Anderson y Elloumi, 2004). Aunque la dimensión tecnológica debe tener también un papel preponderante en estas universidades, -por los beneficios potenciales para la enseñanza a distancia-, la experiencia indica que, paradójicamente, su infraestructura tecnológica no es

[54] Son frecuentes los convenios con empresas y compañías para la financiación de la universidad.
[55] Las universidades abiertas son aquellas que imparten enseñanza a distancia pero no necesariamente on-line, por ejemplo la UNED. Entendemos como universidades *e-learning* aquellas universidades abiertas con enseñanza on-line.
[56] http://www.uoc.edu
[57] http://www.uned.es
[58] http://www.ou.nl
[59] http://www.athabascau.ca/

mayor que en otro tipo de universidades (Guri-Rosenblit, 2005). La arquitectura del CV gira entorno a un modelo de gestión empresarial orientado al tipo de estudiante con dedicación a tiempo parcial. El CV es un portal de servicios en línea para la educación superior: el papel del alumno es el de cliente; el profesor es el profesional-empleado que cumple con los objetivos empresariales; la docencia se organiza en equipos, más que individualmente, donde figura un líder de equipo como responsable. Sus objetivos prioritarios son: la calidad de los cursos, especialmente los materiales didácticos, la flexibilidad para el alumno y la reducción de costes (D'Antony, 2003).

3.2.1.3 Modelo centrado en el estudiante

El objetivo de este modelo es construir un sistema flexible al servicio del alumno, buscando asegurar la continuidad de su formación durante toda su vida profesional (*lifelong learning*).

Este modelo está muy extendido en los países anglosajones y suele darse en universidades con enseñanza mixta (Epper y Garn, 2004; HEFCE, 2005). A diferencia de los anteriores modelos, el CV no es el objetivo de los intereses de la institución, ni existe una preocupación excesiva por el desarrollo de una infraestructura tecnológica específica para el campus. Normalmente, se seleccionan plataformas *e-learning* de carácter comercial o de código abierto robustas y seguras que garanticen el acceso y funcionamiento permanente del catálogo de cursos que ofertan. El CV se concibe como una "factoría de cursos" que se diseñan y construyen de forma centralizada por la universidad. El papel de los profesores se restringe al de tutores-animadores. El alumno dispone de los cursos preparados por equipos de diseñadores de contenidos[60] y cuenta con el apoyo del tutor. Normalmente, además de cursos, el CV integra otros servicios de carácter administrativo[61] en un portal universitario o campus virtual universitario. Los objetivos prioritarios son: (i) la accesibilidad y calidad de los cursos, (ii) el aprendizaje personalizado, y (iii) la fidelización del estudiante.

Este modelo y el modelo centrado en la institución son modelos de tipo industrial o corporativo para la formación profesional. Los cursos o las plantillas suelen presentarse juntos en un modelo mixto institución-estudiante.

[60] Estos equipos están formados por pedagogos e informáticos especialistas en multimedia.
[61] Como el de matriculación, consulta del expediente, servicio de biblioteca etc.

3.2.1.4 Modelo centrado en el profesor

Este modelo tiene como objetivo construir un CV al servicio de las necesidades del profesor. El CV se concibe como una herramienta de apoyo al trabajo del profesor en todas sus facetas, docente, investigadora y de gestión académica, porque se entiende que los profesores son el motor de la actividad del CV y de la universidad. Con esta estrategia, es importante el contacto directo y continuado de los profesores y la institución, por lo que es necesario articular modelos de organización administrativa, con cierto grado de descentralización, que faciliten esta comunicación. Este es el modelo elegido por la UCM en el año 2003 para la construcción del CV (Fernández-Valmayor et. al., 2008).

En este modelo, el CV surge inductivamente a partir de los Espacios Virtuales de Trabajo del Profesor (EVTP). Los EVTP pueden ser (i) espacios de enseñanza y aprendizaje para la docencia de las asignaturas y cursos; (ii) seminarios virtuales de trabajo para apoyar la actividad investigadora y de gestión académica de los profesores; y, finalmente, (iii) espacios virtuales en abierto para la difusión de la actividad del profesor. Son espacios de libre acceso, con una apariencia y funcionalidad semejantes a las páginas web, donde el profesor puede publicar contenidos docentes o de investigación, crear su página personal u organizar eventos científicos fácilmente. En este tipo de CV, la participación del profesor es clave, ya que es el responsable de diseñar, crear y gestionar sus propios espacios virtuales. El papel del alumno está determinado por el profesor. El profesor decide cómo utilizar los recursos que ofrece el CV para facilitar el proceso de aprendizaje del alumno e, incluso, puede hacerle partícipe de la construcción y gestión de los EA (Carabantes, Carrasco y Alves, 2005). Los resultados obtenidos con este modelo en la UCM son muy satisfactorios, como demuestran el incremento constante en el número de inscripciones[62] y el incremento en el número medio de conexiones curso a curso[63].

La diversidad de universidades, sin embargo, hace difícil pensar que exista un único modelo de CV que pueda ser aplicado con éxito en todas las universidades. Igualmente, es difícil definir una tipología completa y precisa de todos los modelos de CV. Lo realmente importante es identificar los valores de las dimensiones del modelo

[62] De 3 500 estudiantes y 26 profesores en el curso 2003-04 a 69283 estudiantes y 4162 profesores en el curso 2007-08. Estas últimas cifras suponen que 7 de cada 8 estudiantes y 2 de cada 3 profesores trabaja en el CV de la UCM. Estas cifras están publicadas en https://www.ucm.es/campusvirtual/CVUCM/

[63] 49 000 conexiones de media al mes en el curso 2003-04 y 2 330 000 en el curso 2006-07 (Fernández-Valmayor et. al., 2008).

para analizar los resultados de un CV. A partir de los modelos analizados, la práctica parece indicar que se obtienen mejores resultados con los modelos que tienden a priorizar la participación de los estudiantes y del profesor –componente didáctica-, frente a los que priorizan los intereses institucionales o los tecnológicos.

3.2.2 La arquitectura de un campus virtual

Desde el punto de vista técnico e informático, un campus virtual universitario puede entenderse como un sistema de información[64] encargado exclusivamente del apoyo a los procesos de enseñanza-aprendizaje e investigación (Britain y Liber, 2004; Fernández-Valmayor et. al., 2008). A nivel universitario, este sistema incluye, como mínimo, un LMS genérico que proporciona las capacidades básicas para crear y gestionar los EA. Además, para adaptar el funcionamiento a los requisitos institucionales y didácticos, se integran otros módulos software[65]: LMS específicos; herramientas satélites cuando el LMS no ofrece determinados recursos; aplicaciones web, bases de datos o herramientas del desarrollador. En la figura 3.8 se muestra una posible arquitectura modular de CV[66] formada por un sistema central –con uno o varios LMS genéricos-, y varios módulos independientes integrados: LMS específicos, portales web educativos, un repositorio de recursos didácticos digitalizados (RDD), y una interfaz de acceso que crea un único entorno web a través del cual los profesores y alumnos ven y entran en sus espacios de aprendizaje.

Desde el punto de vista de los usuarios, los EA pueden tener funcionalidades diferentes –dependiendo de las plataformas o herramientas de soporte-, pero se perciben como parte de un único entorno, el CV.

[64] En informática, un sistema de información es un conjunto de cuatro elementos: usuarios, datos, procesos y herramientas software, cuyo fin es la gestión de datos e información.

[65] La mayoría de las aplicaciones *e-learning* están diseñadas para poder ser ampliadas añadiendo nuevos módulos o reprogramadas o adaptadas parcialmente. Por ejemplo, *Moodle* ofrece un API o librería de pequeños programas para añadir nuevas capacidades o cambiar algunas de sus funciones.

[66] Esta arquitectura está inspirada en la del CV de la UCM (Fernández-Valmayor et. al., 2008).

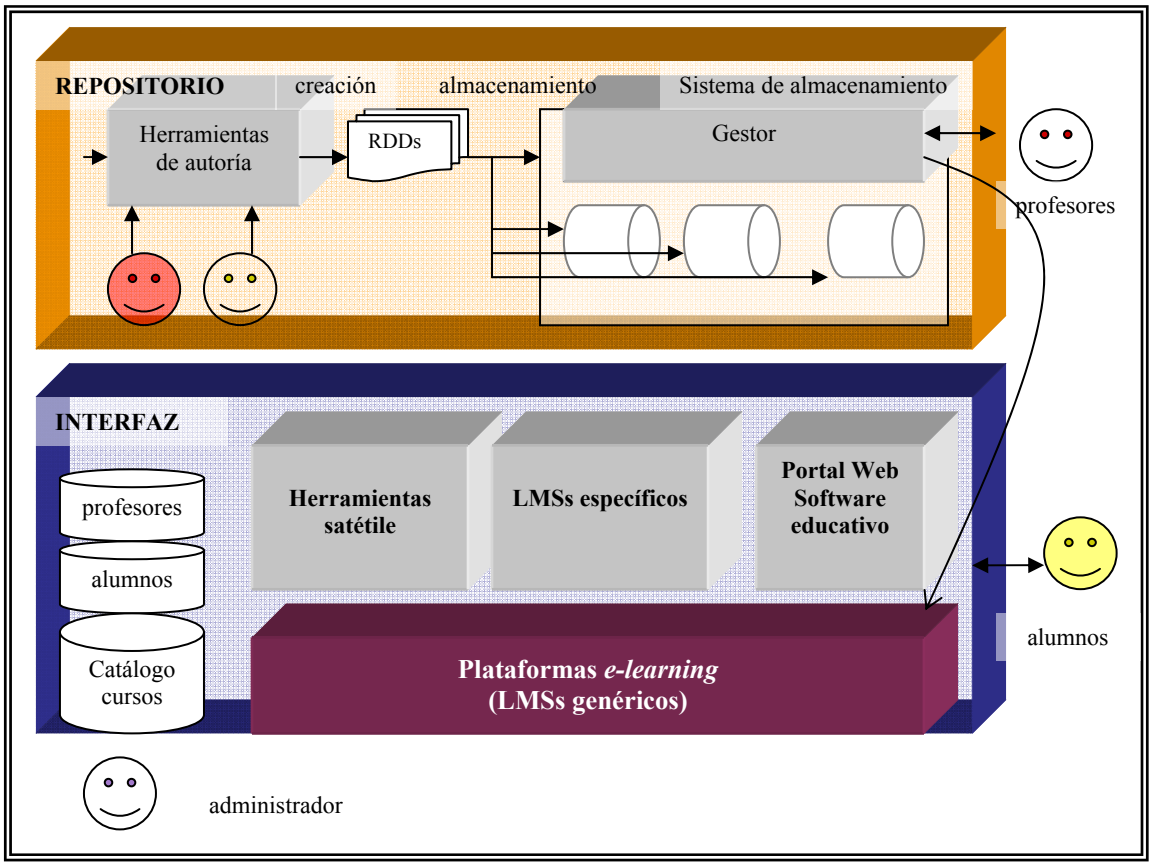


Figura 3.8. Ejemplo de arquitectura modular de un CV

Partiendo de los puntos de vista de los diferentes usuarios[67]- profesor, alumno y administrador-, los servicios básicos que debería ofrecer un CV son los siguientes:

Punto de vista del profesor:

- la creación de sus EA conforme a sus diseños didácticos;
- la gestión interna de estos espacios para su actividad docente; investigadora y de gestión;
- la gestión integrada de todos sus espacios (histórico, altas, bajas, transferencia de datos entre espacios); y
- la incorporación de herramientas complementarias, como los repositorios o los LMS específicos.

Punto de vista del alumno:

- la entrada en los EA en los que está matriculado;
- la utilización de los EA individual y colaborativa durante su aprendizaje; y
- el seguimiento de su rendimiento académico;

[67] En Britain y Liber (2004) puede consultarse una propuesta muy completa sobre los servicios de un CV y una revisión comparativa de los que tienen algunas de las principales plataformas *e-learning*.

- la posibilidad de cambiar al perfil de profesor para ayudar a la construcción de los EA.

Punto de vista del administrador:

- la gestión a nivel de administración de todos los módulos;
- la creación de nuevas aplicaciones y adaptación de las existentes, según las necesidades de los profesores y alumnos;
- la definición de la arquitectura; y
- la definición de los mecanismos de interoperabilidad;
- el control y seguimiento del uso del CV en términos cuantitativos (número de usuarios, número de conexiones, espacio de almacenamiento consumido, tiempos de respuesta de las aplicaciones).

En estas arquitecturas modulares y flexibles de CV es imprescindible la integración de aplicaciones *e-learning* (Santanach et. al, 2007) que sólo es posible si son capaces de interoperar y comunicarse entre ellas. Una de las formas de abordar la interoperabilidad es definir cada uno de los componentes del CV, utilizando un mismo modelo abstracto general. Este modelo está definido en un conjunto de especificaciones estándares[68] de forma que, si todos los módulos software están descritos con las mismas especificaciones, es posible programar la interactuación y el intercambio de información entre ellas. Los programas que resuelven la comunicación entre aplicaciones constituyen otro módulo independiente que, en la figura 3.8, sería el módulo interfaz.

La arquitectura del CV es el soporte real del campus, pero es un soporte complejo. Los CV universitarios actuales son algo más que una simple plataforma *e-learning*, ya que necesitan integrar múltiples herramientas y recursos para conformar las especificaciones institucionales y didácticas. Además, la evolución de esta tecnología es muy rápida y esto repercute en el componente didáctico, porque la docencia y la investigación necesitan tiempo para reflexionar y adaptarse apropiadamente. El éxito o fracaso de un CV depende del uso que los profesores y alumnos hagan de él y, por ello, analizar cómo se está utilizando un CV es imprescindible para el necesario ajuste del componente tecnológico, y, tal vez, el institucional.

[68] IMS Abstract Framework e IMS Tools Interoperability definidas por el IMS Global Learning Consortium. http://www.imsproject.org/

3.2.3. El uso didáctico del campus virtual

El uso didáctico de un CV depende de varios factores: del modelo de CV, de la experiencia del profesor en el CV, del tipo de enseñanza, a distancia o presencial, del área de conocimiento, y de las estrategias pedagógicas utilizadas. En estos momentos, el factor que establece diferencias más palpables respecto de cómo se utiliza un CV es, posiblemente, la experiencia. Por ello, distinguimos los usos didácticos según la experiencia del profesor, y además, suponemos que: (i) el tipo de CV está centrado en el profesor, (ii) la enseñanza es mixta, (iii) la plataforma *e-learning* del CV es neutra respecto a las estrategias pedagógicas, y finalmente (iv), las áreas de conocimiento no van a influir significativamente en el uso global del CV. Estas suposiciones se justifican porque:

(i) El modelo centrado en el profesor es el que permite observar con mayor veracidad cómo se trabaja en el CV; impone, además, menos limitaciones a los profesores para crear y utilizar sus EA, y considera la opinión del profesor un punto básico para evaluar el CV y para tomar decisiones sobre su evolución.

(ii) La enseñanza mixta (*b-learning*) es la modalidad utilizada por la mayor parte de las universidades con un CV y, como ya hemos dicho, es la que, actualmente, está dando mejores resultados en el uso del *e-learning*. Incluso las universidades a distancia que sustituyen la enseñanza presencial por un CV, incluyen en su docencia una mínima presencialidad, lo cual está en consonancia con los resultados de los estudios sobre los tipos de enseñanza, que indican que los estudiantes "a distancia" eligen con preferencia las universidades que les garanticen una presencialidad mínima (Guri-Rosenblit, 2005).

Conviene, además, tener en cuenta que, a pesar de que los CV surgen, a finales de los noventa, con el objetivo de impartir una enseñanza a distancia, los resultados académicos y económicos de este nuevo tipo de entorno no fueron nada satisfactorios (Dondi 2008). Esto motivó un cambio de estrategia hacia la actual enseñanza-aprendizaje mixta, con metodologías que integran lo mejor de los dos entornos: el trabajo en las aulas presenciales con el trabajo en los espacios virtuales. El éxito de esta nueva aproximación ha impulsado el uso del CV en las universidades en estos últimos años (PLS Ramboll, 2004), (Epper y Garn, 2004), (Jenkins, Browne, Walter, 2005).

(iii) Respecto a la neutralidad pedagógica de la plataforma *e-learning*, consideramos que éste es un factor imprescindible para poder observar objetivamente cómo se utilizan diferentes aproximaciones didácticas para enseñar y aprender en un CV. Para dar la posibilidad de tener diversidad didáctica es necesario, como hemos apuntado, considerar un tipo de CV centrado en el profesor, que no limite los procedimientos pedagógicos aplicados en el CV.

(iv) Respecto de la influencia de las áreas de conocimiento en la forma de usar el CV, podemos considerar que, efectivamente, aparecen ciertas diferencias como la preferencia de algunas herramientas según sea la disciplina –por ejemplo el alto uso de test y ejercicios de autoevaluación en Medicina frente a su bajo uso en Ciencias Exactas- o la tendencia hacia estrategias didácticas más reflexivas en Humanidades, y más experimentales en Ciencias (Area, et al., 2008). Sin embargo, desde un punto de vista más global, estas diferencias no condicionan el uso didáctico, ya que el factor que realmente determina de forma significativa cómo se utiliza el CV es la experiencia del profesor en este entorno.

3.2.3.1. El uso didáctico del campus virtual desde la experiencia

El uso didáctico del CV va adaptándose y optimizándose conforme el profesor gana en experiencia. Los EAV pasan de ser paneles de anuncios, o espacios de publicaciones, a ser espacios personalizados y de aprendizaje en colaboración. Esta evolución se realiza en la mayoría de los casos pasando por tres etapas consecutivas que denominamos: a) etapa tecnológica, b) etapa didáctica, y c) etapa de innovación y explotación.

a) La etapa tecnológica es la etapa inicial en la que los profesores se ocupan fundamentalmente de obtener la destreza tecnológica necesaria para dominar el nuevo entorno. Les preocupa no saber utilizar las herramientas o utilizarlas mal con alumnos reales, por lo que es de gran utilidad crearles un EAV de carácter experimental, privado, sin alumnos reales pero con algún alumno ficticio para realizar pruebas. En este tiempo se tiende a utilizar el CV como una página web para la comunicación unidireccional de información académica con sus alumnos y como una herramienta de gestión de sus listas de alumnos (Area, et al., 2008). En general, esta etapa es la más costosa para el profesor en tiempo y esfuerzo. Se puede considerar que, al menos, es necesario un curso académico trabajando con alumnos reales para sentirse cómodo con la plataforma. En esta etapa, la formación y el apoyo al profesor es crítica y debe

ser lo más personalizado posible (Sanz y Fernández-Pampillón, 2009). Sin este apoyo, es fácil fracasar: el profesor no completa esta etapa y abandona. Conviene, en consecuencia, tener en cuenta que en esta fase el profesor se ocupa de la –muchas veces difícil- familiarización con el uso del CV y de llegar a entender cuáles son los beneficios. La introducción en estos nuevos entornos virtuales sin un soporte estable puede conducir al fracaso y abandono del uso del CV.

b) Etapa didáctica. En esta etapa, los profesores ya disponen de la destreza tecnológica necesaria para trabajar rutinariamente en el CV. Las actividades más habituales son: la creación y publicación de materiales didácticos y de información; la tutorización; la discusión; la comunicación permanente con y entre los alumnos; las actividades de los grupos de trabajo; la gestión de alumnos (listas, fichas electrónicas, altas y bajas); la gestión de los trabajos de los alumnos; y el seguimiento y control personalizado de la actividad del alumno.

Desde el punto de vista didáctico, se observa una tendencia a utilizar en el CV las mismas metodologías y estrategias pedagógicas que en la clase presencial, con pequeñas adaptaciones a este medio nuevo cuando son necesarias. En consecuencia, podemos describir los EAV como espacios web –normalmente privados-, más ricos que los paneles "informativos" de la etapa anterior, que contienen materiales didácticos organizados jerárquicamente, enlaces web, trabajos y ejercicios, herramientas de comunicación básicas –foros, correo y, en menor medida, chats-, publicación de calificaciones y recursos para apoyar las actividades en grupo de los alumnos -foros privados, espacios de almacenamiento y, en algunos casos, espacios para la publicación de los trabajos. En algunas áreas de conocimiento, como Ciencias de la Salud y Ciencias Sociales, es habitual el uso de herramientas de evaluación: tests, encuestas y ejercicios de autocorrección. La utilización conservadora del nuevo entorno es, en nuestra opinión, recomendable porque, aunque no explota las potencialidades del CV, ofrece suficientes ventajas como para (i) motivar al profesor en su uso; (ii) apoyar al alumno y facilitarle la realización de algunas de sus actividades; (iii) ahorrar tiempo y esfuerzo al profesor en algunas tareas docentes y de investigación; y (iv) dar la experiencia y seguridad necesarias para que, si lo considera oportuno, pueda innovar sus estrategias docentes. Durante esta etapa, el profesor se acostumbra a utilizar el CV. Este uso habitual es imprescindible para empezar a explorar nuevas formas de enseñar y de aprender. En esta etapa es cuando el profesor puede sentirse motivado por las propuestas de renovación e innovación de

la enseñanza universitaria, tan promocionadas en estos últimos años. El profesor solicita aprender metodologías didácticas y funcionalidades y aplicaciones software nuevas. El apoyo tecnológico, por tanto, sigue siendo necesario, aunque ya no es crítico.

En esta etapa, los profesores destacan como ventajas del CV: la gestión rápida y flexible de los recursos didácticos, de las listas de alumnos, las calificaciones, las prácticas y trabajos, las evaluaciones, y las posibilidades de comunicación síncrona y asíncrona (entre alumnos o con el profesor), el apoyo a las actividades colaborativas, sin las limitaciones físicas y temporales de las sesiones presenciales, y el seguimiento de la actividad del alumno. Existen experiencias en las que los docentes han cuantificado y comparado el rendimiento académico de sus alumnos con el CV y sin el CV, apreciando mejoras en los grupos que aprenden con CV (Fernández-Valmayor, Fernández-Pampillón y Merino, 2007; Fernández-Valmayor, Sanz y Merino, 2008).

Otros profesores, por el contrario, cuestionan la rentabilidad de la substancial inversión en tiempo y recursos que necesitan para "traducir" sus escenarios de la docencia presencial al nuevo medio tecnológico. Esto se corrobora con algunas experiencias desagradables acerca de la caducidad tecnológica de los recursos educativos digitalizados. A pesar de la calidad docente y de la importante inversión de tiempo, esfuerzo y, también, económica, estos recursos didácticos no han podido integrarse en el CV. Otras experiencias negativas provienen de la incompatibilidad entre plataformas, e, incluso, entre versiones de una misma plataforma[69].

c) Etapa de innovación y explotación. Conforme crece la experiencia docente en el CV, el profesor va descubriendo nuevas formas de enseñanza, más rentables, que están basadas en las capacidades del *e-learning*: la accesibilidad y ubicuidad de la información, la facilidad de comunicación, las posibilidades de personalización, la facilidad para compartir conocimientos y recursos, la facilidad para cooperar y colaborar, y la capacidad de integrar en un mismo entorno todos los recursos didácticos. Los profesores, en esta etapa, saben que, con estos nuevos escenarios virtuales, pueden mejorar significativamente la enseñanza y el aprendizaje pero, al mismo tiempo, se encuentran con la necesidad de nuevas funciones en el CV[70]. Es

[69] Ciertos recursos creados en una plataforma no son reutilizables en otras, por ejemplo, los glosarios y las bases de datos de imágenes creadas en WebCT. Incluso los EA (cursos) no puede ser transportados satisfactoriamente entre versiones diferentes de un mismo producto (WebCT 4.0 a WebCT 6.0).

[70] Por ejemplo, incluir funcionalidades de los entornos personales de aprendizaje, plataformas síncronas, y plataformas de gestión de actividades.

precisamente esta necesidad de explotación, más experimentada, la que está provocando un cambio en el *e-learning* hacia lo que algunos autores denominan *e-learning* 2.0 o *i-learning*[71] (integrado, innovador, interpersonal e inclusivo).

En esta etapa los EAV se caracterizan por 1) tener diseños didácticos de carácter socio-constructivistas que promocionan la compartición, la colaboración y cooperación entre alumnos y profesores (López Alonso et al., 2008a; López Alonso et al., 2008b; López Alonso et al., 2009); 2) la promoción de la autonomía del aprendizaje del alumno; y 3) buscar que el estudiante o los grupos de estudiantes personalicen sus entornos de aprendizaje. Estas nuevas características de los EAV coinciden, curiosamente, con las bases didácticas recomendadas para dar continuidad a la enseñanza superior hacia una formación permanente a lo largo de la vida.

La realidad en un CV universitario es que estamos todavía en la segunda etapa y que la innovación y explotación es todavía objeto de investigación y experiencias didácticas menos numerosas. Los profesores usan el CV, mayoritariamente, con los modelos y métodos didácticos de su enseñanza presencial, pero sus resultados son satisfactorios. Los que se introducen en el uso de CV se centran en aprender a utilizar el nuevo entorno y lo utilizan, tímidamente, como espacios de difusión de información y de gestión académica. Es cierto que incorporar el CV a la enseñanza requiere tiempo y esfuerzo pero, también, es indudable que, con un poco de experiencia, el CV ahorra tiempo y esfuerzo y, con un poco más de experiencia, puede incluso ayudar a enseñar y aprender mejor.

3.2.3.2. El uso didáctico y la evolución del *e-learning*

Actualmente, la evolución del uso del CV y de las herramientas *e-learning* está fuertemente influida por la calidad. Existe una necesidad por definir la calidad del *e-learning*, por buscar y definir modelos de CV de calidad, EAV de calidad, plataformas de calidad, recursos educativos de calidad, y métodos y modelos que faciliten el aprendizaje y métodos objetivos para la evaluación de la calidad. Estas cuestiones tienen que ser abordadas de forma coordinada en los tres ámbitos de actuación tecnológico, institucional y docente. En el ámbito tecnológico destacan las siguientes cuestiones: la definición de metodologías y herramientas para crear y mantener recursos educativos interoperables que resuelvan el problema de la

[71] En Dondi (2008).

caducidad tecnológica; el diseño y desarrollo de plataformas más potentes y flexibles capaces de interoperar entre ellas y con otros componentes software; la capacidad de crear y reutilizar fácilmente los recursos didácticos; y la definición de directrices para la construcción de EAV más eficaces.

Entre las cuestiones relativas a la gestión institucional subrayamos:

(i) el reconocimiento y la valoración del trabajo del profesor en el CV. Es necesario reconocer que aunque la dedicación al CV es muy productiva en términos académicos, para el profesor, es una carga más en su actividad habitual que, de no reconocerse adecuadamente, supondrá un lastre en su promoción profesional que actualmente depende de su actividad investigadora; y

(ii) la valoración de los EVA y los materiales docentes electrónicos como parte de la producción científica y didáctica del profesor. Realmente en la actualidad no se valoran los EVA, a pesar de que en muchos casos son verdaderos "libros electrónicos" de gran interés pedagógico; tampoco se valora de igual manera un manual didáctico en papel que un manual en formato electrónico. Afortunadamente, se está empezando a trabajar intensamente en la búsqueda y definición de estándares de calidad para el *e-learning* y en el reconocimiento oficial de la actividad "virtual" de los profesores (Proyecto EFQUEL – the European Foundation for Quality in eLearning)[72]. Sin embargo, conviene tener en cuenta que sólo disponemos de, aproximadamente, una década de experiencia en CV universitarios, lo que nos obliga a preguntarnos (i) si es posible, en este corto espacio de tiempo, poder extraer conclusiones sobre la calidad; y (ii) si tenemos perspectiva y datos suficientes como para evaluar y valorar con objetividad las aportaciones del *e-learning* a la enseñanza y el aprendizaje. Y mientras… ¿cómo van las instituciones a valorar el trabajo virtual de los profesores?[73]

Finalmente, de las cuestiones por resolver en el ámbito docente destacamos: la definición de modelos cognitivos eficaces para aprender en entornos virtuales, la definición de métodos de construcción y explotación de EVA y de recursos didácticos

[72] http://www.qualityfoundation.org/

[73] Actualmente, en España la Agencia Nacional de Evaluación de la Calidad y Acreditación (ANECA) no dispone de criterios para valorar la actividad en *e-learning* de los profesores. En el documento (academina_faq02_080314, 2008) sólo se encuentra la siguiente referencia a este tipo de actividad docente "[…] ¿Cómo se justifica el material docente on-line que se encuentra a disposición de los alumnos a través del aula virtual, accesible sólo para los alumnos matriculados en las asignaturas correspondientes?
No existe una regla fija para la justificación de este tipo de material. El solicitante debe presentar lo que considere más conveniente, siempre que no suponga un exceso de documentación […]".

durables, accesibles y de calidad y la definición de métodos de evaluación de la rentabilidad didáctica de los EVA y los recursos didácticos digitalizados utilizados para la enseñanza y el aprendizaje.

Probablemente, trabajando desde éstas últimas cuestiones de ámbito didáctico, que sólo pueden ser respondidas por cada profesor con su conocimiento sobre la disciplina que imparte, su comprensión de las plataformas *e-learning* y su experiencia y saber docente, podamos avanzar adecuadamente para resolver las cuestiones de los ámbitos tecnológicos e institucional.

3.3. La aportación de los tesauros en el contexto del *e-learning*

En la etapa más avanzada de uso didáctico de los entornos *e-learning,* que hemos denominado de innovación y explotación didáctica", los profesores y/o autores de los materiales didácticos incorporan el uso tesauros. En el contexto del *e-learning*, los tesauros y, en general los vocabularios, aportan un conjunto de términos y sus significados, procedentes del lenguaje natural o de un lenguaje controlado, para: 1) definir e identificar el contenido de los recursos educativos, también llamados objetos de aprendizaje, bien mediante su directa asociación con el recurso o bien aportando los valores a las propiedades de los metadatos de los recursos; y 2) representar los conceptos relacionados semánticamente acerca de una disciplina o área de conocimiento al que pertenecen los recursos educativos. La idea básica es que sirvan para comprender el ámbito representado, con el objeto de facilitar el aprendizaje, la búsqueda y selección de los términos y/o recursos más apropiados para expresar una idea o resolver una consulta.

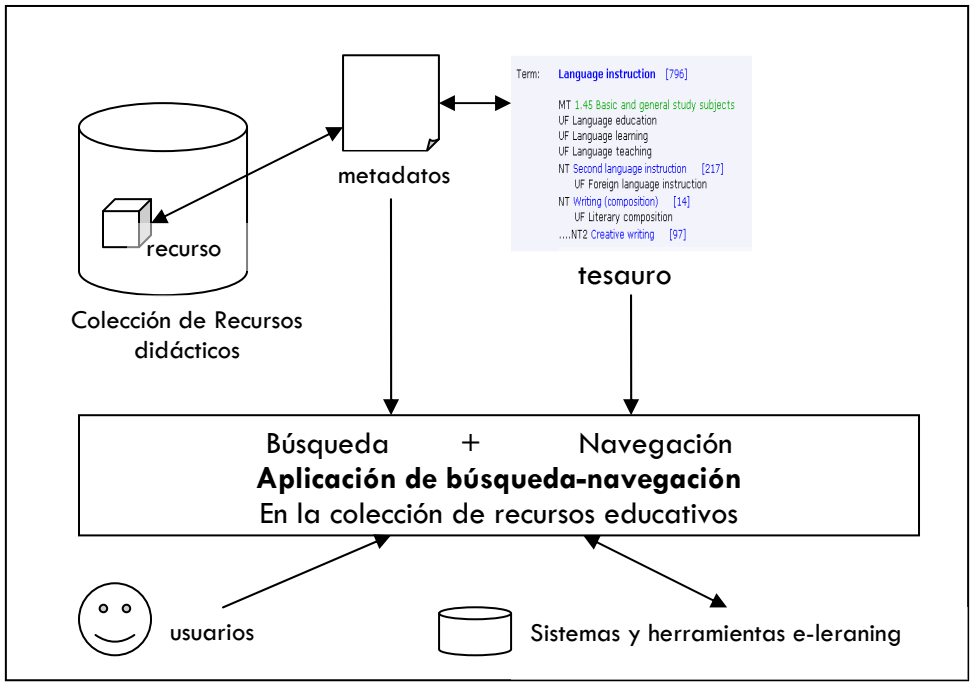


Figura 3.9. Uso de tesauros para la descripción de colecciones de recursos educativos

Uno de los usos de los vocabularios más habituales en el *e-learning* es la descripción del contenido de los recursos educativos digitalizados en los metadatos que documentan estos recursos (ver capítulo 2, sección 2.5). Los vocabularios aportan los términos precisos para definir las propiedades de los recursos. Para ello, estos vocabularios deben estar declarados en los metadatos y ser accesibles vía Web de forma que los sistemas de búsqueda y navegación de recursos puedan utilizarlos.

Además, los modelos de metadatos estándares recomiendan el uso de vocabularios *generales* con el objetivo de facilitar la compartición y reusabilidad de los recursos educativos con independencia de dónde, cuándo y quién lo creo. El vocabulario es el sistema de referencia que las herramientas y sistemas *e-learning* utilizan para interpretar y gestionar los recursos educativos (figura 3.9).

En el contexto del *e-learning*, el modelo actual de metadatos estándar más difundido es el Learning Object Model, LOM (IEEE-LOM 1484.12.1, 2002). En este modelo recomienda que, para incrementar la interoperabilidad, se utilicen una serie de vocabularios de referencia, públicos[74], para dar valores a algunas de las propiedades. Estas recomendaciones se repiten, también, en otros estándares de metadatos no específicamente educativos, como el Dublin Core (Dublin Core, 2008).

[74] Algunos de estos vocabularios forman parte del propio modelo LOM.

El modelo de metadatos LOM contiene 78 propiedades organizadas en nueve grupos; cada uno describe un aspecto del recurso educativo: general, ciclo de vida, metametadatos, técnico, educativo, propiedad intelectual, relaciones con otros objetos, anotaciones y clasificación (figura 3.10). Cada grupo, a su vez, contiene otras propiedades más específicas, para refinar la descripción, hasta un máximo de cuatro niveles. Las propiedades del último nivel contienen valores simples que pueden ser libremente definidos por el usuario, por ejemplo *general.title*, o estar sujetos a una gramática patrón, por ejemplo *general.identifier* o pertenecer a un vocabulario controlado (tabla 3.1).

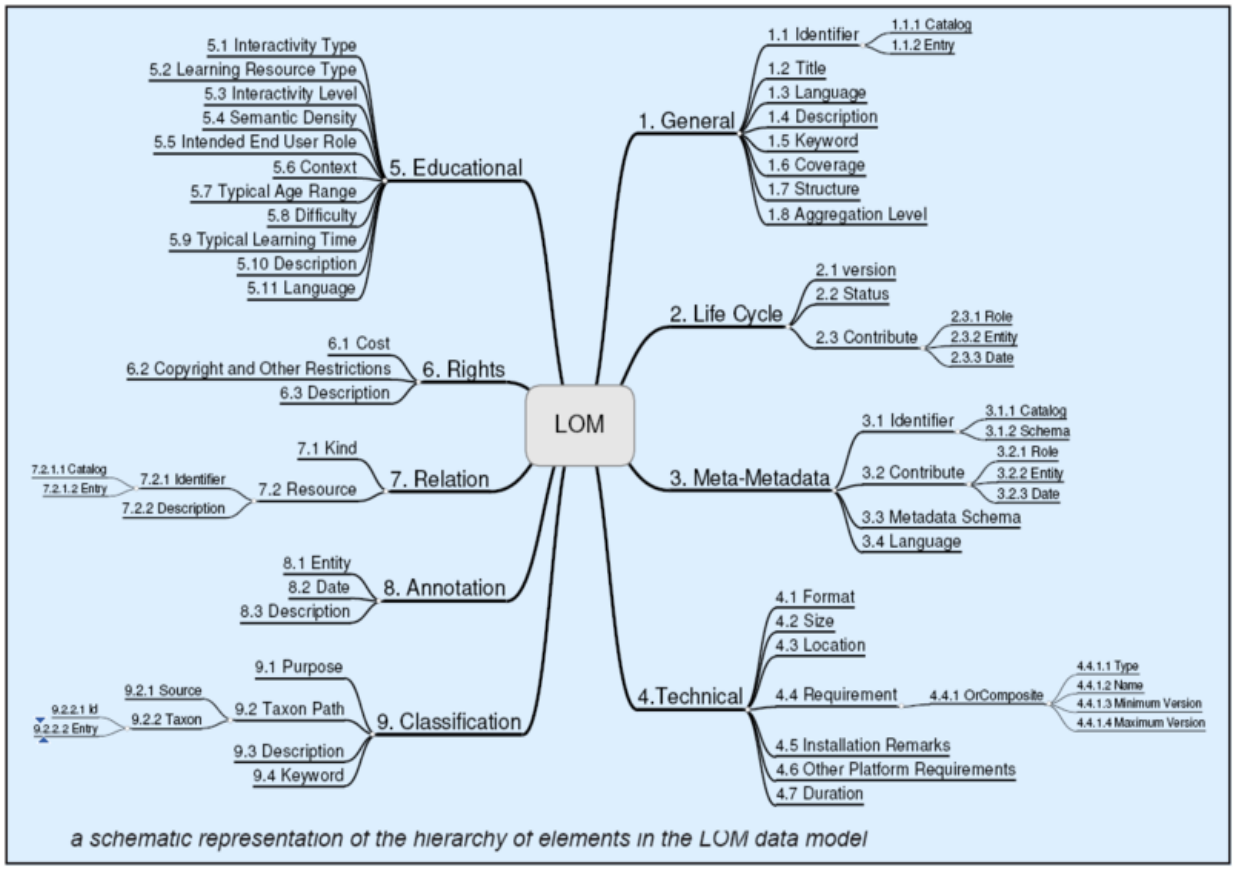


Figura 3.10. Estructura de LOM (fuente (Barker, 2005))

Los vocabularios recomendados por LOM deben ser vocabularios autorizados, lo que significa que deben tener una mínima difusión, estabilidad, aplicabilidad, un formato estándar y estar comúnmente acordados por algún comité de expertos (IMS Meta-data, 2004). No se recomienda el uso de vocabularios propietarios porque merma la capacidad de interoperabilidad. Sin embargo, las experiencias de uso de LOM confirman que las aplicaciones existentes no siguen esta recomendación (CEN CWA 14871, 2003).

Elemento LOM	Vocabularios recomendados por LOM
general.language	**Vocabulario simple** El estándar ISO (código de lenguas) ISO 639:1988 ISO3166 http://www.iso.ch/
general.keywords classification.purpose (discipline) [75]	**Clasificaciones** LCC (Library of Congress Classification) http://lcweb.loc.gov LCSH (Library of Congress Subject Headings) http://lcweb.loc.gov DDC (Dewey Decimal Classification) http://www.oclc.org/oclc/fp/ UDC (Universal Decimal Classification) http://zeus.slais.ucl.ac.uk/udc/ GEM (http://www.thegateway.org/about/documentation/gem-controlled-vocabularies/vocabulary-subject **Taxonomías o tesauros** Sistema de Clasificación de la ACM (Ciencias de la Computación) http://www.acm.org/class/1998/homepage.html Cyc General Financial Taxonomy Suite http://taxonomies.cyc.com/ Medical Subject Headings (MESH) http://www.nlm.nih.gov/mesh/

[75] Un buen almacén de vocabularios temáticos puede consultarse en http://www.taxonomywarehouse.com. En español se propone el tesauro ETB MEC CCAA (Berrocal et al., 2008), también Otra buena recopilación de tesauros de referencia, en diversas lenguas, puede encontrarse en Alonso, (2001).

	Art and Architecture Thesaurus http://www.getty.edu/research/conducting_research/vocabularies/aat/
technical.format	**Vocabulario simple** MIME -RFC 2046 http://www.mhonarc.org/~ehood/MIME/
educational.learningcontext	**Clasificación** GEM (Gateway to Educational Materials) Controlled Vocabularies[76] http://www.thegateway.org/about/documentation/gem-controlled-vocabularies/ Cyc Education Taxonomy Suite http://taxonomies.cyc.com/
technical.requirement.Name	Lista de valores http://merlot.org/search/AdvArtifactSearch.po
general.aggregationlevel	No hay recomendación
educational.learningresource type	**Vocabulario simple** DCMI Type Vocabulary http://dublincore.org/documents/dcmi-type-vocabulary GEM http://www.thegateway.org/about/documentation/gem-controlled-vocabularies/
educational.interactivitytype	GEM http://www.thegateway.org/about/documentation/gem-controlled-vocabularies/
educational. interactivitylevel	No hay recomendación
educational.intendedenduser role	GEM http://www.thegateway.org/about/documentation/gem-

[76] GEM son nueve vocabularios, uno por cada elemento raíz. Para describir la disciplina o tema se tomaría el vocabulario correspondiente a "Subject Element": http://www.thegateway.org/about/documentation/gem-controlled-vocabularies/vocabulary-subject

	controlled-vocabularies/
educational.difficulty	No hay recomendación

Tabla 3.1. Algunos vocabulario, de tipo de datos vocabulario, recomendados para los elementos LOM (IMS Metada-data, 2004)

En LOM los vocabularios constituyen descripciones universales para los aspectos *técnicos* y semánticos de los recursos educativos (CEN CWA 14871, 2003). El tipo de vocabularios para la descripción de las *propiedades técnicas* son vocabularios simples o listas de no más de diecisiete valores, lo que obliga a utilizar términos con una semántica muy general. Los vocabularios utilizados pueden ser (ver tabla 3.1):

- Vocabularios basados en estándares internacionales, como el código de países (ISO 3166, 1999) o el código de lenguas (ISO 639,1988).
- Vocabularios autorizados externos a LOM, como el tesauro ETB en el repositorio AGREGA (Berrocal et al., 2008); y
- Vocabularios LOM que forman parte del modelo LOM (tabla 3.2).

No.	*Name*	*Comments on vocabularies*
1.7	*Structure*	atomic, collection, networked, hierarchical, linear.
1.8	*Aggregation level*	1-4
2.2	*Status*	draft, final, revised, unavailable
2.3.1	*Role*	author, publisher, unknown, initiator, terminator, validator, editor, graphical designer, technical implementer, content provider, technical validator, educational validator, script writer, instructional designer, subject matter expert.
3.2.1	*Role*	creator, validator.
4.4.1.1	*Type*	operating system, browser.
4.4.2	*Name*	pc-dos, ms windows, macos, unix, multi-os, none or netscape communicator, ms-internet explorer, opera, amaya.
5.1	*Interactivity Type*	active, expositive, mixed.
5.2	*Learning Resource Type*	exercise, simulation, questionnaire, diagram, figure, graph, index, slide, table, narrative text, exam, experiment, problem statement, self assessment, lecture.

Tabla 3.2. Vocabularios LOM: términos del vocabulario en tercera columna y las propiedades LOM se especifican en la primera y segunda columnas (CEN CWA 14871, 2003)

Los vocabularios recomendados para *la clasificación* de los recursos son las categorías y taxonomías y los tesauros, que incluyen relaciones semánticas de hiperonimia e hiponimia (figura 3.11).

Figura 3.11. Uso de clasificaciones-taxonomías en los metadatos LOM (ver Browse in Categories). También se utilizan vocabularios del tipo simple, por ejemplo, en la propiedad Material Type y texto en lenguaje natural, como en Description.

Para obtener mayor eficacia en la recuperación de recursos, se utilizan tesauros en vez de clasificaciones y taxonomías. Los tesauros aportan mayor riqueza semántica que las clasificaciones y taxonomías para poder localizar términos equivalentes y términos relacionados semánticamente con otros. En el siguiente capítulo, se examina, con mayor detalle, la ganancia en eficacia en las búsquedas basadas en tesauros.

El uso de ontologías no es frecuente en LOM, probablemente porque las ontologías son más difíciles de construir y procesar que los tesauros y taxonomías, y además tienen como propósito fundamental la definición del conocimiento de un dominio, a diferencia de LOM, que tiene como objetivo la indexación, búsqueda y navegación en las colecciones de recursos educativos.

3.3.1 Un ejemplo de clasificación de recursos educativos con metadatos LOM y taxonomías o tesauros

El problema que presenta el modelo de metadatos LOM es que es demasiado amplio y complejo, por lo que es costoso y difícil de aplicar. En la práctica, el modelo se simplifica y se utiliza una parte de sus propiedades (Friesen, 2004): sólo las que son más útiles para localizar cada recurso en una colección. Una de estas propiedades útiles es la clasificación. Esta propiedad tiene el objetivo de clasificar el contenido del recurso respecto de taxonomías y/o tesauros recomendados, proporcionando una descripción semántica útil para búsquedas más inteligentes que las basadas en valores de las propiedades[77] Los vocabularios habitualmente utilizados son las clasificaciones/taxonomías, como *Library of Congress Classification* de Estados Unidos (LCC), *Dewey Decimal Classification* (DDC) y *Universal Decimal Classification* (UDC), y los tesauros, como el *European Treasury Browser* (ETB) para el ámbito de la educación, *Art & Architecture Thesaurus* (AAT) para arquitectura y arte o *Medical Subject Headings* (MESH) para medicina.

En esta sección se presenta un ejemplo de cómo utilizar, siguiendo las recomendaciones del estándar LOM, un vocabulario de tipo clasificación-taxonomías, para describir recursos utilizando la propiedad clasificación (IMS Meta-data, 2004). El propósito es, por un lado, mostrar, con una aplicación concreta, las ideas revisadas en este capítulo y en el capítulo anterior sobre el uso de vocabularios para la explotación didáctica; por otro lado, poner de manifiesto lo difícil que es, para los profesores que son especialistas en su disciplina pero no en metadatos y tesauros, utilizar vocabularios y metadatos para documentar recursos educativos.

La propiedad clasificación de LOM tiene como propósito la clasificación sistemática del contenido de los recursos educativos respecto de uno o varios vocabularios. Para poder usar un vocabulario es necesario que, como mínimo, tenga asociado un identificador URI[78] para poder referenciarlo. Además, es recomendable, pero no necesario, que:

1) la estructura y el contenido del vocabulario estén definidos en un formato procesable (txt, XML, …) y estándar[79]; y

[77] Las taxonomías, tesauros y ontologías permiten ampliar o precisar las consultas (ver capítulo 4, sección 4.2).

[78] RFC 3986 URI Generic Syntax 2005. Accesible en: http://www.ietf.org/rfc/rfc3986.txt
Identificación y direccionamiento de recursos W3C, accesible en: http://www.w3.org/Addressing/

[79] (ANSI/NISO Z39.19, 2005), (IMS VDEX Model, 2004). El capítulo siguiente revisa los estándares para la construcción de taxonomías y tesauros.

2) esté accesible en la Web y haya sido publicado y registrado en registros públicos y reconocidos[80].

La propiedad clasificación puede utilizarse una o varias veces para uno o varios propósitos, lo que permite tener un recurso educativo clasificado según diferentes aspectos. Por ejemplo, si se desea clasificar un recurso según la disciplina a la que pertenece su contenido se utilizará una clasificación con el propósito de "disciplina" y se utilizará un vocabulario que contenga una clasificación de las disciplinas, como la clasificación del repositorio Merlot[81] (figura 3.11). También, es posible clasificarlo respecto del contenido del recurso educativo, propósito, idea, utilizando, por ejemplo, el tesauro multilingüe europeo ETB[82].

Cada clasificación está formada por cuatro subpropiedades (figura 3.12) que indican (1) el *propósito* (purpose) de la clasificación, uno o varios (2) *caminos taxonómicos* (taxonpath), términos relacionados por hiperonimia e hiponimia, una (3) *descripción*(description) en lenguaje natural y las (4) *palabras clave* (key words).

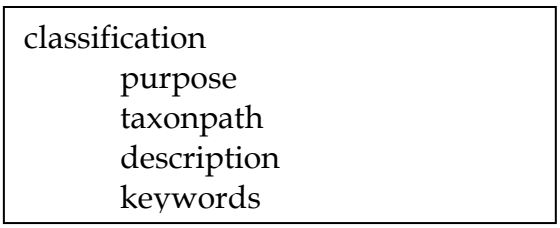


Figura 3.12. Estructura de la propiedad clasificación

La subpropiedad (1) *propósito* contiene un valor simple escogido de un conjunto de valores prefijados por LOM (discipline, idea, prerequisite, educational objetive, accessibility restrictions, educational level, skill level, security level). Se refiere al propósito u objetivo del vocabulario que debe coincidir con el propósito con el que se clasifica el recurso educativo. Por ejemplo, si el propósito tiene como valor "disciplina" significa que este vocabulario concreto describe las disciplinas, o materias de un área de conocimiento determinado y que se utiliza para describir la disciplina, o disciplinas, a la que pertenece el recurso.

La propiedad (2) *camino taxonómico* contiene un camino[83] de una clasificación o taxonomía que sirve para clasificar el recurso. Se especifica definiendo las propiedades

[80] (Hillmann, et al., 2006), ISO/IEC 11179 Metadata Registries [http://metadata-standards.org/11179/]
[81] http://www.merlot.org/merlot/index.htm
[82] http://etb.eun.org/eun.org2/eun/en/etb/content.cfm?lang=en&ov=7208
[83] Conjunto de nodos conectados desde la raíz de la jerarquía.

opcionales *fuente (source)* y *taxón (taxon)*[84]. La fuente sirve para indicar el responsable, organización, persona, etc., del vocabulario, taxonomía o tesauro. El taxón es cada uno de los "nodos" del camino taxonómico, y están formados, a su vez, por un identificador y una entrada. La entrada contiene un término del vocabulario. El identificador es la referencia alfanumérica única que tiene el término en el vocabulario.

Para describir un recurso se pueden utilizar múltiples caminos taxonómicos que describen diversos caminos taxonómicos, bien dentro de la misma taxonomía si el recurso pertenece a varias disciplinas, o bien respecto de varias taxonomías, para incrementar su interoperabilidad. Supongamos que necesitamos clasificar el recurso titulado "Animated Technical Dictionary" del repositorio Merlot (figura 3.11), como un recurso de la disciplina Ciencias y Tecnología (SCIENCE AND TECHNOLOGY), y dentro de esta disciplina en el subárea de Tecnologías de la Información (INFORMATION TECHNOLOGY) y, más concretamente, en el tema de "Redes" (NETWORKING). Para ello se define el siguiente camino taxonómico utilizando la clasificación de disciplinas del repositorio[85]:

```
taxonpath
  source: Merlot categories
  taxon
    id: B
    entry: SCIENCE AND TECHNOLOGY
    taxon
      id: F
      entry: INFORMATION TECHNOLOGY
      taxon
        id: 180
        entry: NETWORKING
```

Pero también podría estar clasificado respecto del tesauro ETB, versión española ETB-LRE MEC-CCAA V.1.0 (Berrocal et al., 2008), utilizando otro camino taxonómico. Esto significa que el recurso puede ser interpretado respecto de las categorías de Merlot y también del tesauro ETB-LRE MEC-CCAA v.1.0.

[84] El taxón es cada una de las subdivisiones de la clasificación, desde la raíz hasta el término final elegido para describir el objeto.

[85] El identificador del camino taxonómico se obtiene concatenando cada uno de los identificadores de los taxones que lo componen. En el ejemplo sería: BF 180.

taxonpath
source: ETB-LRE MEC-CCAA V.1.0
taxon
id: M70.50
entry: Ciencias aplicadas / tecnologías / técnicas
taxon
id:
entry: tecnología
taxon
id:
entry: tecnologías de la información y la comunicación
taxon
id:
entry: redes de comunicación de datos

Finalmente, las clasificaciones pueden contener una *descripción* y *palabras clave*. En la primera, la descripción, se puede incluir una descripción en lenguaje natural sobre el recurso (ver etiqueta Description en figura 3.11); en la segunda, las palabras clave, se puede incluir el conjunto de palabras clave que caracterizan al recurso y que no tienen que proceder necesariamente de un vocabulario. En definitiva, la clasificación completa correspondiente al recurso del ejemplo de la figura 3.11, tienen el propósito de indicar la disciplina a la que pertenece dicho recurso y está formado por: i) un solo camino taxonómico con términos del vocabulario Merlot, y ii) una pequeña descripción. El código LOM es el siguiente:

classification
purpose: discipline
taxonpath
source: Merlot categories
taxon
id: B
entry: SCIENCE AND TECHNOLOGY
taxon
id: F
entry: INFORMATION TECHNOLOGY

taxon

id: 180

entry: NETWORKING

description: Animated technical dictionary.

Para que esta descripción se pueda procesar automáticamente se escribe en un archivo de texto con un lenguaje de marcado XML propuesto, también, en el estándar LOM:

```
<classification>
<purpose>
        <source uniqueElementName="source">LOM-ESv1.0</source>
        <value uniqueElementName="value">discipline</value>
</purpose>
<taxonPath>
        <source>
        <string language="en">Merlot categories</string>
        </source>
        <taxon>
        <id>B</id>
        <entry>
        <string language="en"> SCIENCE AND TECHNOLOGY</string>
        </entry>
        </taxon>

        <taxon>
        <id>F</id>
        <entry>
        <string language="en"> INFORMATION TECHNOLOGY </string>
        </entry>
        </taxon>

        <taxon>
        <id>180</id>
        <entry>
        <string language="es">NETWORKING</string>
        </entry>
        </taxon>

</taxonPath>
<description>
<string language="en"> Animated technical dictionary </string>
</description>
</classification>
```

También se podría ampliar esta clasificación refiriéndola a otros vocabularios, como por ejemplo del tesauro ETB, incluyendo un camino taxonómico más procedente de este vocabulario, con lo que la descripción LOM-XML quedaría:

```
<classification>
<purpose>
        <source uniqueElementName="source">LOM-ESv1.0</source>
        <value uniqueElementName="value">discipline</value>
</purpose>
<taxonPath>
        <source>
```

```
        <string language="en">Merlot categories</string>
        </source>
        <taxon>
        <id>B</id>
        <entry>
        <string language="en"> SCIENCE AND TECHNOLOGY</string>
        </entry>
        </taxon>

        <taxon>
        <id>F</id>
        <entry>
        <string language="en"> INFORMATION TECHNOLOGY </string>
        </entry>
        </taxon>

        <taxon>
        <id>180</id>
        <entry>
        <string language="es">NETWORKING</string>
        </entry>
        </taxon>

</taxonPath>
<taxonPath>
        <source>
        <string language="es">ETB-LRE MEC-CCAA V.1.0</string>
        </source>

        <taxon>
        <id>M70.50</id>
        <entry>
        <string language="es">Ciencias aplicadas / tecnologías / técnicas </string>
        </entry>
        </taxon>
        <taxon>
        <entry>
        <string language="es">tecnología</string>
        </entry>
        </taxon>
        <taxon>
        <entry>
        <string language="es">tecnologías de la información y las comunicaciones</string>
        </entry>
        </taxon>
        <taxon>
        <entry>
        <string language="es">redes de comunicación de datos</string>
        </entry>
        </taxon>

</taxonPath>
<description>
<string language="en"> Animated technical dictionary </string>
</description>
```

Este ejemplo muestra cómo se construye una descripción de un recurso basada en vocabularios y metadatos estándares. Estas descripciones facilitan la localización y acceso automático a dicho recurso, así como la durabilidad e independencia de las aplicaciones que lo procesen. Sin embargo, exige que las personas que crean los

metadatos tengan un conocimiento amplio en metadatos LOM y en vocabularios, lo que supone un esfuerzo considerable para la mayor parte de los profesores que, en general, desconocen los detalles de los estándares de metadatos. Simplemente para poder construir cada clasificación del ejemplo previo, la persona que documenta debe conocer en primer lugar, la sintaxis y semántica de la propiedad clasificación; en segundo lugar, la estructura y el contenido de los vocabularios, y en tercer lugar, la materia o dominio de conocimiento de los recursos a fin de poder seleccionar el término o términos más apropiados de un tesauro de referencia. Este inconveniente limita el uso de metadatos y vocabularios en aquellos contextos académicos en los que los profesores, expertos en sus disciplinas, son los responsables de crear, documentar y utilizar los recursos educativos. Sólo en los centros e instituciones educativas con modelos de CV centrados en la institución o en el estudiante, y que cuentan con suficientes recursos económicos y de personal, es posible crear colecciones de recursos educativos documentadas con metadatos y vocabularios conforme los estándares (Friesen, 2004; Heath, et.al, 2005; Hepp, 2007; Caceres, 2007).

3.4. Resumen y conclusiones del capítulo

> "La tecnología puede ser el catalizador para la enseñanza y el aprendizaje si se utiliza de forma que promueva la reflexión, discusión y colaboración en la resolución de tareas o problemas" (Murria, 1999).

En este momento, podemos afirmar que las plataformas *e-learning* son un catalizador tecnológico para la enseñanza y el aprendizaje universitario. Permiten crear espacios de aprendizaje (EA) en Internet, con una amplia gama de funcionalidades al servicio de distintos tipos de enseñanza y aprendizaje. Son sistemas o aplicaciones software, principalmente LMS, orientados a la creación y gestión de múltiples EA con diferentes tipos de usuarios. También pueden considerarse plataformas *e-learning* otros sistemas más específicos en su orientación pedagógica o funcional, como los CMS, los LCMS, los EPA, los LAMS y los sistemas de aprendizaje síncronos. Estos sistemas específicos pueden operar independientemente o integrados en LMS para así extender sus capacidades. Además, existen otras herramientas satélites para realizar algunas funciones no resueltas en las plataformas *e-learning*.

Este entramado tecnológico es complejo de entender, usar y mantener por el personal no informático, como son los profesores o equipos de profesores y, por eso, sólo es posible utilizar *e-learning* si se dispone de campus virtuales. Los campus virtuales

son espacios en Internet compuestos por todos los EA de una institución, que es la responsable de su diseño, implantación y mantenimiento. Normalmente, se construyen integrando una o varias plataformas *e-learning* generales y/o específicas, en arquitecturas normalmente modulares y flexibles donde la interoperabilidad es primordial. Los campus virtuales universitarios facilitan el uso de tecnología *e-learning* a los profesores y alumnos con el objetivo de mejorar su trabajo académico, la calidad de su enseñanza-aprendizaje, de optimizar recursos y, en definitiva, de poder ser una institución de enseñanza superior más competitiva.

Se pueden considerar distintos tipos o modelos de CV, según primen los objetivos institucionales, tecnológicos o didácticos. El modelo didáctico centrado en el profesor es el más flexible y el que ofrece más posibilidades docentes y discentes. Es habitual que se utilice en la modalidad de enseñanza mixta. El uso didáctico del CV depende, fundamentalmente, de la experiencia que tiene el profesor en este entorno. Este uso puede clasificarse en tres etapas: tecnológica, didáctica, y de innovación y explotación. La primera etapa la denominamos tecnológica porque el profesor se encuentra determinado por el grado de conocimiento y destreza en el manejo de las plataformas *e-learning* del CV. Es crítico, en este primer momento, garantizarle orientación y soporte personalizado. En una segunda etapa, didáctica, el profesor ya incorpora muchas de las posibilidades del CV a su enseñanza presencial, y obtiene, en general, resultados satisfactorios. En la tercera etapa, innovación y explotación, la experiencia del profesor le permite cambiar sus métodos de enseñanza para adaptarlos a un mejor aprovechamiento de las posibilidades del CV. Aunque en el momento actual se puede considerar que el uso del CV está en una segunda etapa, didáctica, los profesores comienzan a experimentar nuevas aproximaciones didácticas y herramientas que permitan avanzar en el uso rentable y de calidad de la tecnología *e-learning*.

En este momento, una de las herramientas que necesitan los profesores y que se está ya incorporando a los CV son los repositorios de recursos educativos. Estas herramientas incorporan tesauros electrónicos para facilitar la comprensión del contenido del repositorio. Los tesauros ofrecen una descripción con términos del lenguaje natural o del lenguaje de especialidad de las disciplinas o áreas de conocimiento al que pertenecen los recursos educativos del repositorio. Al mismo tiempo, estos términos se utilizan para describir, en los metadatos, el contenido de cada uno de los recursos del repositorio. De esta forma los tesauros sirven para ayudar

al usuario a encontrar recursos didácticos en repositorios con gran cantidad de recursos educativos. Sin embargo, existen dos inconvenientes que limitan la efectividad de los tesauros, y metadatos, como herramientas de descripción y búsqueda en contextos *e-learning*: i) no disponer de tesauros que describan con precisión y de forma comprensible para el usuario los contenidos de los recursos educativos de los repositorios y ii) el esfuerzo que supone documentar los recursos educativos con tesauros cuya terminología y estructura, normalmente compleja, es desconocida.

Con el objetivo de construir tesauros más eficaces para estos entornos *e-learning* que faciliten la documentación y acceso a los recursos y contenidos didácticos, se revisarán en los próximos dos capítulos, los modelos de construcción de tesauros analizando si permiten resolver las limitaciones, indicadas en el párrafo anterior, y las necesidades específicas, definidas en las conclusiones del capítulo segundo, de los tesauros *académicos de explotación* en entornos *e-learning*.

Capítulo 4

El modelo de los estándares de construcción de tesauros de explotación

Una característica general de los tesauros es su estructura compleja, basada en relaciones y agregaciones de términos. Para construirlos se ha consensuado y consolidado a lo largo de estos últimos cuarenta años, un modelo general que describe, estructura y visualiza su contenido de naturaleza altamente relacional. Este modelo está orientado a un tipo de tesauros denominado *tesauros de explotación*. Se llama tesauro de explotación a aquel que está orientado a mejorar la efectividad de los sistemas de almacenamiento, recuperación y navegación Web (ASNI/NISO Z39.19, 2005).

En este capítulo se revisa, fundamentalmente, este modelo general de los estándares para la construcción de tesauros de explotación siguiendo la siguiente estructura: en la primera sección, se hace una introducción a los modelos de datos que son objeto de este capítulo y del siguiente; en la segunda sección, se especifican las características y los requisitos de los tesauro de explotación; en la tercera sección, se presenta el modelo estándar para la construcción de tesauros monolingües de explotación; en la cuarta sección, se analiza la aplicación del modelo estándar al diseño de los esquemas de datos tradicionales, alfabético y sistemático, de los tesauros de explotación; finalmente, en la quinta sección, se presenta un resumen del capítulo y unas conclusiones sobre la aplicación de los requisitos y del modelo estándar revisados al caso de los tesauros académicos de explotación.

4.1. Introducción a los modelo de datos

En Informática, un *modelo de datos* es una herramienta para diseñar y describir la organización formal de la estructura, tipos de contenidos, restricciones y operaciones de un conjunto de datos, de información o el conocimiento[1] relativo a un domino u

[1] Salton y McGill definen de forma precisa, y acertada, los conceptos de dato, información y conocimiento en Informática de la forma siguiente: los datos son los hechos almacenados físicamente; la información es la interpretación de los datos hecha por máquinas o personas; y el conocimiento es la información incorporada y almacenada en estructuras de conocimiento susceptibles de ser consultadas y procesadas (Salton y McGill, 1986).

organización[2] (Bertino et al., 2001). Los modelos de datos se utilizan, principalmente, para construir bases de datos y bases de conocimiento.

Para diseñar los tesauros se han utilizado, sin embargo, modelos y metodologías no informáticos, provenientes de la Lexicografía[3] y, sobre todo, de Biblioteconomía y Documentación (Taylor, 2006). En estos últimos años, con la progresiva digitalización de los tesauros y su aplicación en los sistemas de RI, de gestión de bases de datos y de gestión de recursos digitalizados, se empieza a incorporar el uso de los modelos de datos informáticos, pero combinándolos con los modelos tradicionales y estándares para la construcción de los tesauros y vocabularios en general (Jones, 1993; CEN CWA 14871, 2003; Aitchison y Clarke, 2004).

Los modelos de construcción de tesauros estándares, como veremos, organizan el contenido del tesauro en la presentación, de forma que se facilite la lectura de su contenido. Por el contrario, los modelos de datos informáticos se centran en representar la estructura y funciones del tesauro, de forma independiente de la presentación. Esta separación entre estructura y presentación aporta numerosas ventajas entre las que se destacan que: i) permite resolver la duplicación de información y los problemas de inconsistencia en las relaciones, ii) reduce los costes de mantenimiento, iii) permite el procesamiento automático del contenido, iv) facilita la interoperabilidad entre tesauros, y v) permite crear múltiples presentaciones, actualizadas, de un contenido (Taghva et. al, 1999; Gibbon, 2000; CEN CWA, 2005).

Los elementos conceptuales que proporcionan los modelos de datos sirven para definir la estructura, las restricciones del dominio que hay que modelar y, opcionalmente, el conjunto de operaciones para la gestión, actualizaciones y consultas, de los datos (Miguel y Piattini, 1997). El resultado de aplicar un modelo de datos a un dominio es un *esquema de datos*. El esquema de datos incluye la definición formal de la organización lógica y las restricciones de los objetos o entidades, por ejemplo, términos y categorías, del dominio modelado[4]. Así, por ejemplo, el modelo gráfico Entidad-Relación (Chen, 1976), utiliza los elementos: entidades, propiedades y relaciones[5] para obtener un esquema de datos gráfico (figura 4.1).

---

[2] No todos los modelos de datos informáticos son completos, en el sentido de que no siempre son capaces de definir, además de la estructura del sistema, su funcionamiento.

[3] (Roget, 1852)

[4] Este esquema teórico o *esquema de datos* se puede definir en términos lingüísticos como la interpretación de un vocabulario respecto de un modelo de datos (Gibbon, 2000).

[5] Una *entidad* representa un objeto o concepto del dominio que hay que modelar, por ejemplo un término; un *atributo* representa alguna propiedad de una entidad, por ejemplo, si es descriptor, la nota de ámbito o

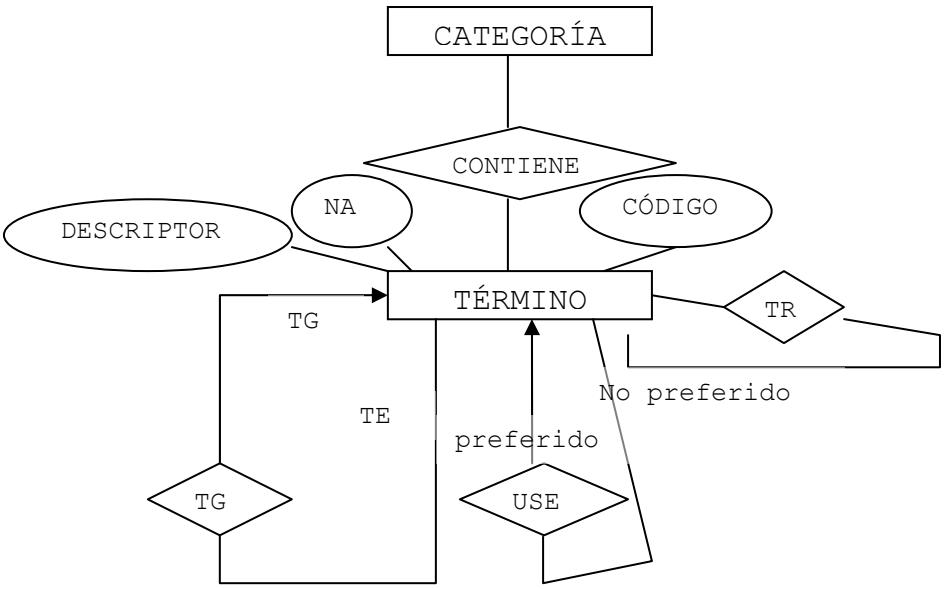


Figura 4.1. Ejemplo de diagrama Entidad Relación para modelar un tesauro[6].

Los modelos de datos se pueden clasificar según su grado de abstracción en modelos conceptuales, modelos de implementación de datos, y modelos físicos (Ullman, 1998).

Los *modelos conceptuales* o de alto nivel, utilizan conceptos y lenguajes cercanos a los esquemas de organización mental humana y, por tanto, fácilmente comprensibles e interpretables para las personas, pero que no son, o difícilmente son, intepretables directamente por las máquinas[7]. Necesitan ser traducidos a un modelo de datos más cercanos a las máquinas, aunque tienen la ventaja de ofrecer una vista global de todo el dominio. La figura 4.1 muestra un esquema conceptual que hemos creado con el modelo de datos conceptual Entidad-Relación. En el capítulo siguiente se revisarán otros modelos de datos conceptuales como las redes semánticas o el modelo UML.

Los *modelos de implementación de datos* tienen menos capacidad de abstracción y son más cercanos a las estructuras de almacenamiento y organización de datos utilizadas por los sistemas informáticos. Como ejemplos se pueden citar los modelos relacional, que revisaremos, también en el próximo capítulo, jerárquico y en red.

---

el código; un vínculo o relación describe una asociación entre dos o más entidades, por ejemplo, término genérico, use o término relacionado.

[6] Las etiquetas TG, TE, USE, TR, NA se corresponden con la notación estándar "Término Genérico", Término Específico, USE (reenvios a término preferido en relaciones de equivalencia), "Término Relacionado" y "Nota de Ámbito" (UNE 50106, 1990).

[7] Se necesitan herramientas especiales para transformar estos esquemas en esquemas procesables. Por ejemplo, algunas herramientas CASE permiten diseñar a nivel conceptual y automáticamente transforman los esquemas de datos conceptuales en esquemas de datos de implementación.

Finalmente, los *modelos de datos físicos o de bajo nivel*, no abstraen las representaciones de los datos sino que describen exactamente cómo son las estructuras de almacenamiento de los datos en el sistema informático. Estas estructuras son un conjunto de archivos de datos y de índices, que dependen del hardware y del sistema operativo de la máquina.

Cuando se diseña un esquema de datos, normalmente, se comienza utilizando un modelo de datos conceptual, porque permite representar más fácilmente la estructura y requisitos del dominio que se modela. Después, el esquema de datos conceptual se traduce, utilizando un modelo de implementación de datos, a un esquema de implementación de datos. Finalmente, este esquema de implementación se traduce al esquema físico que, en definitiva, es el conjunto de archivos con datos interrelacionados que se almacena en el sistema informático. Este proceso de traducción, entre esquemas de datos de niveles de abstracción diferentes, se denomina *correspondencia* o *mapping.* En la actualidad, se realiza, automáticamente, bien con herramientas CASE de diseño asistido por ordenador, o bien con los Sistema de Gestión de la Base de Datos (SGBDs)[8].

El uso de modelos de datos aporta tres ventajas principales: 1) la independencia lógica de los datos, 2) la independencia entre los esquemas de datos y su presentación, y 3) la posibilidad de utilizar software de gestión de carácter general para manipular automáticamente cualquier base de datos. La *independencia lógica de los datos* se refiere a la capacidad para realizar modificaciones en el esquema de datos de un nivel de abstracción, sin tener que modificar, o modificando de forma controlada, los esquemas de datos de los otros niveles y los programas de aplicación. La *independencia entre los esquemas de datos y la presentación* implica poder mostrar al usuario sólo la parte de la base de datos, o la información que se desee y de la forma que se desee sin tener que alterar el esquema de datos. En entornos multiusuario, por ejemplo, es posible definir diferentes presentaciones según el tipo de usuario. Otra de las ventajas de utilizar modelos de datos es que facilita la interoperabilidad entre bases de datos y entre SGBD.

Los modelos de datos pueden proporcionar, además de un conjunto de estructuras para organizar los datos, un conjunto de operadores para la consulta y manipulación formal de estas estructuras (Bertino et al., 2001). Tanto el esquema de datos como sus operaciones se expresan mediante lenguajes propios del modelo o definidos como

[8] Un SGBDs es una aplicación de propósito general para la gestión (definir, crear, mantener, consultar y controlar el acceso) de la base de datos utilizando el esquema de datos

extensiones del mismo. Los *Lenguajes de Definición de Datos (DDL)* sirven para definir el esquema de los datos mientras que los *Lenguajes de Gestión de Datos (DML)*, también llamados lenguajes de consulta en RI, definen las operaciones con datos.

Las principales aplicaciones de los modelos de datos para la construcción de tesauros se encuentran en la RI, en los sistemas de Representación de Conocimiento (RC), y en la Web Semántica, para representar el contenido o semántica de los recursos digitales. Para elegir el modelo de datos adecuado se deben tener en cuenta las características y requisitos del tesauro y de la aplicación o del entorno en el que se va integrar. Antes de revisar el uso de los modelos de datos para la construcción de tesauros se examinan las características y requisitos de los tesauros orientados a la explotación.

4.2. Características y requisitos de los tesauros de explotación

4.2.1. Características

La primera característica de un tesauro es que, en general, sirve para representar las relaciones semánticas entre las palabras en una lengua, tesauro monolingüe, o de las palabras de varias lenguas, tesauro multilingüe. Este rasgo distintivo es el que debe prevalecer en la selección de un modelo de datos, sea cual sea el propósito del tesauro.

La segunda característica es que el contenido se debe organizar en dos niveles complementarios (Aitchison et. al., 2000; Lancaster,1986). El primer nivel es el de la microestructura[9] que describe cada término con sus relaciones directas (figura 4.2). El segundo nivel es la macroestructura[10] y se corresponde con la organización global del tesauro, normalmente en agrupaciones temáticas de jerarquías de términos y/o agrupaciones de términos equivalentes y/o de términos relacionados (figura 2.26).

[9] Organización interna de cada artículo lexicográfico.
[10] Organización del léxico recogido en el vocabulario, el tesauro en este caso.

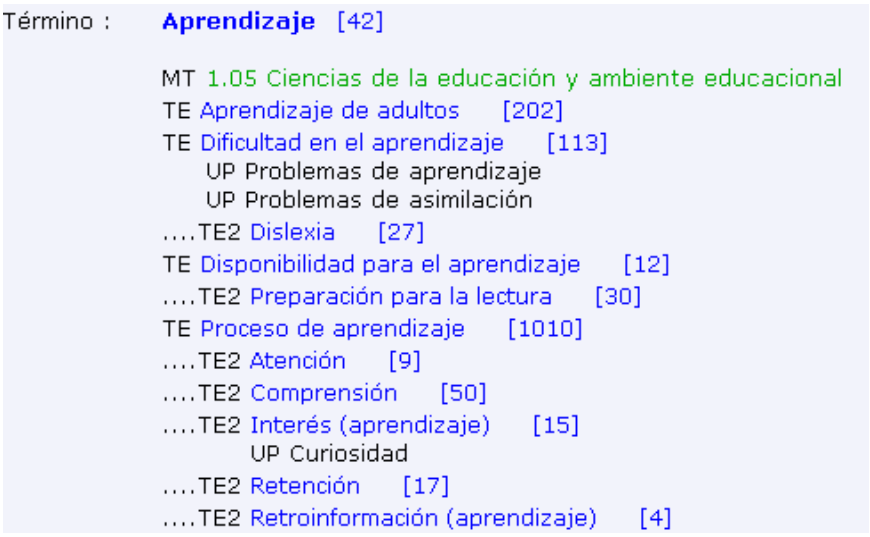


Figura 4.2. Microestructura del término Aprendizaje

Estos niveles son claves para definir las formas de presentación y acceso al contenido de un tesauro. La presentación basada en la microestructura muestra el tesauro como una lista alfabética de entradas (figura 4.3). Cada entrada es una microestructura formada por el término y todas sus relaciones con otros términos. La presentación basada en la macroestructura, además, muestra la estructura reticular del tesauro, normalmente basada en jerarquías. Las presentaciones más completas integran la microestructura en la macroestructura, de forma que la organización global de la obra resulta de la macroestructura, pero en cada término se incluye su microestructura, por ejemplo el tesauro ETB, (2008). Los modelos estándares de construcción de tesauros definen con precisión ambos niveles estructurales, sus presentaciones y accesos, como veremos en el próximo epígrafe.

Figura 4.3. Presentación alfabética del Tesauro de Derecho del CSIC[11].

La tercera característica de los tesauros, especialmente de explotación, es el tipo de contenido. Los términos y las relaciones semánticas de equivalencia, asociativas y jerárquicas constituyen el contenido básico de un tesauro monolingüe. Además el tesauro puede incluir categorías que agrupan a los términos. En los tesauros multilingües se añaden los términos relacionados en las diferentes lenguas. Este tipo de contenido está definido en los estándares para la construcción de tesauros, que revisaremos con detalle en la siguiente sección. En los tesauros de explotación, además, se incluyen como un elemento más los objetos de contenido indexados por el tesauro (ANSI/NISO Z39.19, 2005).

La cuarta característica de los tesauros se refiere a la gestión, es decir, cuáles son los modos de consulta y los métodos de mantenimiento. En los tesauros están establecidos los modos de presentación, acceso y modificación del contenido, a partir de los cuales se definen las operaciones consulta que son básicamente obtener[12] 1) la microestructura de un término, 2) la jerarquía de términos a partir de uno dado, 3) las jerarquías globales del tesauro, 4) las categorías y sus términos y 5) subcategorías si existen.

[11] Cada término es un acceso a su microestructura. La barra superior muestra las presentaciones posibles: jerárquica y alfabética. Disponible en: http://thes.cindoc.csic.es/index_DEREC_esp.html

[12] En la siguiente sección se describen con detalle los modelos de presentación estándares y estas operaciones de consulta basadas en dichos modelos.

Respecto al mantenimiento, también existen una serie de reglas bien establecidas que indican cuándo y cómo hacer una modificación de inserción, actualización o borrado (Aitchison et al., 2000; Lancaster, 1986). Estas reglas y los modos de presentación deben definirse en esquema de datos del tesauro.

4.2.2. Requisitos

Los requisitos para el diseño y desarrollo de un tesauro de explotación se pueden clasificar en tres tipos: 1) los propios del tesauro, 2) los relativos al dominio de conocimiento y 3) los relativos a su uso.

1) Los requisitos, propios de un tesauro, son (ANSI/NISO Z39.19, 2005; Lancaster, 1986):

(i) La eliminación de la ambigüedad. Teniendo en cuenta que cada concepto debe estar descrito con una única forma lingüística, y si existieran varias formas deben controlarse y regularizarse agrupándose (mediante la relación de equivalencia) para evitar múltiples puntos de acceso para un mismo concepto. La polisemia se resuelve con las notas de ámbito, que son breves descripciones que precisan el significado, o con un término entre paréntesis que se añade para cualificar al término polisémico (figura 4.4)

(ii) El control de la sinonimia, para evitar fallos debidos a que los términos de consulta, cercanos semánticamente al término elegido para un concepto, no se correspondan con el concepto o no aparezcan en el tesauro.

(iii) La necesidad de establecer el tipo de relación apropiado entre los términos, para ayudar al agente, sea una persona o una aplicación informática, que indexa la información o recurso y para ayudar al usuario, persona o aplicación, a encontrar el término más apropiado para el objeto u objetos que hay que describir o buscar. Y

(iv) La necesidad de que los términos estén comprobados y validados como los mejores para representar un concepto. Esta comprobación/validación se realiza desde tres aspectos: literario, del usuario y de la organización. El primero, literario, asegura que para un concepto determinado el término incluido en el tesauro es el más utilizado en el dominio (colección de recursos, documentos, información o textos de referencia). La segunda, del usuario, incluye el estudio de los términos utilizados por los usuarios para referirse a un concepto determinado en sus consultas. Finalmente, el aspecto referente a la organización, comprueba

que los términos seleccionados son los preferidos por la organización u organizaciones que van a utilizar el tesauro.

2) Los requisitos propios del dominio de conocimiento que hay que tener en cuenta, son (Gibbon, 2000):

(i) La cobertura, que es el tamaño del tesauro, tanto desde el aspecto extensivo, número de términos, como intensivo, número de campos de información en cada término, y que está limitado al conjunto de objetos que tienen que describirse y clasificarse.

(ii) El alcance, que es el dominio de conocimiento del tesauro. Cuando está delimitado, los tesauros se denominan técnicos, de especialidad o de dominio específico, como en el caso que nos concierne en este trabajo de investigación con tesauros para la explotación de recursos educativos digitalizados. Y

(iii) El tipo de usuarios, si son expertos del dominio o usuarios de carácter general. Normalmente los usuarios no son los creadores de los tesauros, sin embargo en entornos académicos los creadores y destinatarios coinciden son los profesores. Los alumnos son usuarios aunque, en menor medida, pueden participar en la construcción.

3) Finalmente, los requisitos específicos de los tesauros orientados a la explotación de contenidos de información o recursos digitalizados son: expresividad, flexibilidad, efectividad y accesibilidad:

(i) La expresividad se refiere a la capacidad que debe tener el tesauro de describir de forma comprensiva el dominio de su alcance. La estructura de referencias cruzadas del tesauro debe ofrecer una ayuda positiva para que el usuario seleccione los términos más apropiados a su necesidad y, en el caso de una búsqueda exhaustiva, debe conducir al usuario a todos los términos que podrían ser relevantes (Lancaster, 1986).

(ii) La flexibilidad se refiere a la capacidad de asumir e incorporar fácilmente los cambios y actualizaciones que surgen a nivel informático y lingüístico (Calzolari, 1991).

(iii) La efectividad de un tesauro de explotación se mide utilizando los dos indicadores que también miden la efectividad de un sistema de RI: la precisión y la exhaustividad (Lancaster, 1986). Como se explicó en el capítulo 2, sección 2.4, la efectividad del tesauro depende de la especificidad de sus términos y de los el grado en el que se han tenido en cuenta los requisitos, antes mencionados, de

eliminación de la ambigüedad, control de sinonimia, exactitud en la definición de relaciones y la utilización de términos contrastados. Un tesauro muy específico permite obtener una mayor precisión en la descripción de la información, pero al mismo tiempo esta precisión complica la búsqueda porque es necesario un conocimiento del dominio y del lenguaje de especialidad profundo para poder expresar adecuadamente la consulta. Por el contrario, si el tesauro es general el usuario tiene más probabilidades de encontrar, utilizando conceptos muy generales, los documentos o recursos que busca, pero al mismo tiempo puede encontrarse con una gran cantidad de documentos no deseados (figura 2.4). Y

(iv) La accesibilidad es la facilidad con la que los usuarios consultan el tesauro (Lancaster, 1986). En las especificaciones estándares para la construcción de tesauros monolingües (ANSI/NISO Z39.19, 2005) el acceso al tesauro está establecido con cuatro formas de acceso o presentación: alfabética, jerárquica, sistemática y gráfica. Tomándolas como referencia se definirán, en la próxima sección, las operaciones básicas que un modelo de datos debe incluir para garantizar estos los modos estándar de acceso al tesauro.

En resumen, el modelo de datos que se utilice para diseñar un tesauro debe ser capaz de recoger las características del tesauro siguientes: 1) la estructura altamente relacional organizada en macroestructura y microestructura; 2) el contenido formado por términos, relaciones y categorías; 3) los cambios permanentes, tanto en estructura como en contenido; 4) los modos de acceso al contenido que están establecidos en los estándares; además, debe tener en cuenta los requisitos 5) propios de un tesauro: evitar la ambigüedad, controlar la sinonimia, representar con exactitud las relaciones y utilizar los términos más adecuados; 6) del dominio: cobertura, alcance y tipo de usuario; y 7) los relativos al tipo de aplicación: expresividad, flexibilidad, efectividad y accesibilidad.

4.3. Los modelos de datos estándar para la construcción de tesauros de explotación: el estándar ANSI-NISO Z39.19

La construcción de tesauros de explotación no es una actividad nueva. La experiencia de más de medio siglo[13] se ha ido recogiendo y normalizando en guías, recomendaciones y

[13] El primer tesauro de este tipo se construyó para el Engineering Information Center of E.I. Dupont de Nemurs en el año 1959 (Holm y Rasmussen, 1961).

estándares desde los años 60[14]. En la actualidad, se puede considerar que existe un modelo de descripción y visualización de tesauros consensuado y consolidado propuesto en los estándares más destacados para la construcción de tesauros monolingües: ANSI/NISO Z39.19, de EEUU, el ISO 2788, internacional y promovido por la UNESCO, (ISO 2788, 1986), su versión española (UNE 50106, 1990)[15], y el estándar británico BS-5723[16]. Estos estándares han ido evolucionando e influyéndose mutuamente desde sus primeras versiones en el año 1974. En esta sección presentamos una síntesis de todos ellos con especial énfasis en la última versión del ANSI/NISO Z39.19 del año 2005. La razón es que esta versión: (1) recoge las propuestas de los otros estándares; (2) está orientado a los tesauros de explotación, objetivo de este trabajo de tesis, como se puede comprobar en la sección de introducción: "…la descripción consistente de objetos de contenido para facilitar la recuperación en los sistemas de Información (SI) y de organización del conocimiento (KOS)…"; y (3) considera un rango amplio de vocabularios controlados puesto que incluye, además de los tesauros, las listas de términos, los anillos de sinónimos y las taxonomías.

Es importante destacar que los estándares para la construcción de tesauros monolingües no son modelos de datos, sino que proporcionan una especificación de estos sistemas léxicos en un marco bien establecido, unificado y consensuado. Incluso aun cuando el estándar ANSI Z39.19 incluye la definición de cuatro tipos de estructuras, correspondientes a los cuatro tipos de vocabularios controlados, listas, anillos, taxonomías y tesauros, la realidad es que para los tesauros, sólo proporciona recomendaciones sobre la organización de la presentación del contenido[17]. Esta revisión se ha organizado en tres partes: 1) contenido, que revisa los elementos que constituyen un tesauro, 2) acceso al contenido o presentación, y 3) modos de modificación del contenido.

[14] Una de las primeras guías para la construcción de tesauros fue COSATI (1967). Recoge la experiencia en la construcción del Thesaurus of Engineering and Scientific Terms llevada a cabo por el Committee on Scientific and Technical Information, un comité mixto entre el Departamento de Defensa y el sector industrial de EEUU. La primera versión (1974) de los estándares ANSI/NISO Z39.19 y ISO 2788 pueden considerarse derivados de la guía COSATI (Lancaster, 1986).

[15] "Directrices para el establecimiento y desarrollo de tesauros monolingües", desarrollada por el *Grupo de trabajo sobre Vocabularios estructurados para la recuperación de información* del Comité 50 de AENOR.

[16] La primera versión de BS 5723, de 1979, ya integra una metodología de construcción de tesauros basada en facetas con las metodologías basadas en campos temáticos del ISO 2788 y ANSI Z39.19.

[17] Ver sección *11.1.2. Determine the Structure and Displey formats* del ANSI/NISO Z39.19-2005

4.3.1. El contenido del tesauro

Los elementos estándar que contiene un tesauro son los siguientes:

1. Términos
2. Categorías (etiquetas de nodos)
3. Relaciones semánticas
4. Otros:
    - Objetos de contenido
    - Índices

4.3.1.1. Términos

Un término es una unidad léxica, una o más palabras, que representa un único concepto (figura 4.4). Los términos pertenecen al lenguaje natural y su selección y formato se realiza de acuerdo a unas reglas o principios bien establecidos que se pueden consultar en los estándares[18]. Cuando es necesario se añaden a los términos propiedades específicas que los describan con más detalle:

(i) las notas de ámbito (NA), son descripciones en lenguaje natural para precisar el significado del término en el tesauro;

(ii) las notas históricas (NH), son informaciones que se añaden para mantener un historial del término, como, por ejemplo, la fecha de inclusión y modificaciones;

- **ANAFORA (LINGÜISTICA)**
- N.A.:
  - V. A. EL SUBENC.-ANAFORA BAJO LENGUAS Y GRUPOS DE LENGUAS
- T.R.:
  - GRAMATICA COMPARADA Y GENERAL
- U.P.:
  - REFERENCIAS CRUZADAS (LINGÜISTICA)

Figura 4.4. Término ANAFORA[19] (tesauro de la Biblioteca UCM)

Cuando existen varios términos para referirse a un concepto el *término preferido* o *descriptor* es el término elegido para representar un concepto y son los únicos que tienen asignado un objeto de contenido. Esta elección se realiza aplicando un conjunto

[18] Por ejemplo, los sustantivos que expresan conceptos abstractos deben expresarse en singular (conductivismo, lingüística, arquitectura, etc.); los sustantivos contables en plural (libros); y los incontables en singular (nieve, agua). (UNE 50106, 1990).

[19] Este término se desambigua mediante el cualificador (LINGÜÍSTICA). También se precisa el significado mediante una nota de ámbito (NA).

de reglas, que pueden ser las propuestas por los estándares o bien diseñadas por los autores del tesauro. En cualquier caso, estas reglas deben quedar bien definidas en los preliminares del tesauro. El resto de los términos, sinónimos o variantes léxicas, son *términos no preferidos* y se llaman *términos entrada* porque son punteros al término preferido. Los términos preferidos forman el vocabulario controlado de indexación, mientras que los no preferidos constituyen un vocabulario controlado de entrada al índice (figura 4.5).

La existencia de términos candidatos, provisiones o borrados es opcional y depende de la metodología de mantenimiento definida en el tesauro. Indica los posibles estados de "aceptación" de un término para su inclusión en el tesauro.

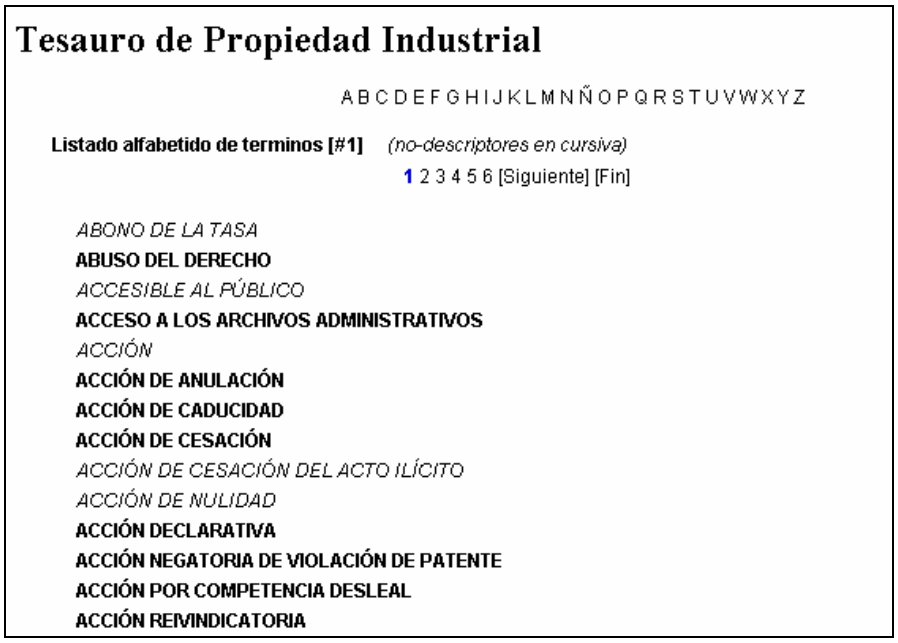


Figura 4.5. Términos descriptores (en negrita) y no descriptores (en cursiva).[20]

4.3.1.2. Categorías o etiquetas de nodo

La categoría es un grupo de términos asociados por criterios semánticos o estadísticos normalmente pertenecientes a un mismo nivel jerárquico. Por ejemplo, en la figura 4.6 se muestran los hipónimos del término 'paintings' organizados en tres categorías llamadas 'paintings by form', 'by location or context', 'by material or technique'. Sirven para precisar el significado de los términos, aportando información sobre las características que comparten un conjunto de ellos. En las jerarquías de términos hacen explícitos los criterios de especificidad de los términos de niveles inferiores. Las categorías pueden ser de dos tipos (Aitchison et al., 2000): 1) temáticas o por campos,

[20] Fuente: Tesauro de Propiedad industrial del CSIC, disponible en: http://thes.cindoc.csic.es/index_PROIND_esp.html

cuando se refieren a clases, tipos de dominio, disciplina o área temática, y 2) facetas, cuando se refieren a cada uno de los aspectos o dimensiones de un dominio, disciplina o área temática[21]. Normalmente, cuando se diseña un tesauro primero se clasifican los términos del tesauro en categorías principales y, después, se subdivide cada categoría principal en facetas mediante un análisis facetado[22]. Por ejemplo, para organizar los términos de un tesauro de Arqueología podrían definirse categorías de tipo clasificación: Etnología, Arqueología, Reproducciones y Material documental, y las facetas: Área cultural, periodo cultural, lugar, tema y tipo.

Figura 4.6. Uso de categorías para clasificar los términos[23]

4.3.1.3. Relaciones semánticas.

Las relaciones semánticas entre términos a las que ya nos hemos referido en el capítulo 2, sección 2.3, se explicitan mediante enlaces bidireccionales, -siempre debe existir el recíproco- por lo que pueden considerarse operadores binarios entre términos y/o categorías:

[21] Véase la sección 4.4.2.
[22] Véanse los métodos de construcción de tesauros en el capítulo 7.
[23] En este caso se clasifican los hipónimos del término **paintings** según sus características de forma, localización o contexto y materiales o técnica. Fuente: tesauro ATT, disponible en: http://www.getty.edu/research/conducting_research/vocabularies/aat/#sample

Forma infija: término A ENLACE término B

Forma prefija: ENLACE (término A, término B)

Los tipos básicos de relaciones de los tesauros son: la equivalencia, las de jerarquía y la de asociación. Cada tipo de relación tiene una marca o abreviación estándar, excepto la relación jerárquica de inclusión[24] (tabla 4.1)

Relación	**Marca**	Abreviación
	Inglés (español)	Inglés (español)
Equivalencia	USE (USE)	U (U)
	USED FOR (USE PARA)	UF (UP)
Jerarquía	BROADER TERM (TÉRMINO GÉNÉRICO)	BT (TG)
	NARROWEB TERM (TÉRMINO ESPECÍFICO)	NT (TE)
Asociación	RELATED TERM (TÉRMINO RELACIONADO)	RT (TR)

Tabla 4.1. Relaciones semánticas básicas y convenios de marcado estándares

1) *Relación de tipo equivalencia*. Se corresponde con la relación semántica de sinonimia y cuasi-sinonimia, pero también con las relaciones léxicas de variantes léxicas, envío genérico y referencias cruzadas a cada palabra de un término compuesto. Es una relación asimétrica entre un término elegido como el término preferido y el término o términos no preferidos.

Se marca con las etiquetas U(Término_nopreferido,Término_preferido) ó USE(Término_nopreferido,Término_preferido).

La relación inversa es

UF(Término_preferido,Término_nopreferido) ó

USED FOR(Término_preferido,Término_nopreferido)

2) *Relación de tipo jerarquía*. Se corresponde con las relaciones semánticas de hipero-hiponimia, todo-parte, pertenencia (instancia) o inclusión en una categoría. Por ejemplo, en la figura 4.7, los términos INDUSTRIA SIDERÚRGICA, Acería, Industria del Acero, Industria Siderometalúrgica mantienen una relación de equivalencia, y además, todos ellos tienen como término genérico (TG) el término SECTOR INDUSTRIAL.

En el estándar se incluye la posibilidad de distinguir, mediante marcas más refinadas, los tres tipos de relaciones jerárquicas:

[24] A partir de aquí introduciremos las marcas en español (UNE 50106,1990) y en inglés (ANSI/NISO Z39.19, 2005).

- hiper/hiponimia: BTG/NTG (Broader Term Generic/Narrower Term Generic)
- parte/todo: BTP/NTP (Broader Term Partitive/Narrower Term Partitive)
- pertenencia: BTI/NTI (Broader Term Instance/Narrower Term instance)

Figura 4.7. Relaciones de equivalencia (sinonimia) entre los términos Acería e Industria siderúrgica[25]

3) *Relación de tipo asociativa*. Es una relación simétrica que conecta términos entre los que existe una "intersección de significados" (figura 4.8), y que no son equivalentes ni existe entre ellos relaciones jerárquicas (figura 4.9).

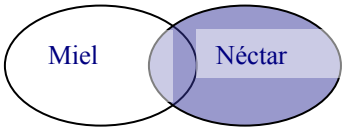


Figura 4.8 Intersección de significados entre los términos miel y néctar

[25] Fuente: Tesauro de Urbanismo del CSIC, disponible en: http://thes.cindoc.csic.es/index_URBA_esp.html)

EN : Honey	BT : Productos de la colmena
FR : Miel	RT : Abeja melífera
ES : Miel	RT : Néctar
AR : عسل	RT : Azúcar
ZH : 蜂蜜	UF : Miel de abeja
PT : Mel de abelha	UF+ : Aquamiel
CS : med	
JA : 蜂蜜	
TH : น้ำผึ้ง	
SK : med	

Figura 4.9. Términos relacionados (asociados) con miel (RT) en el tesauro multilingüe AGROVOC[26]

No es fácil determinar con objetividad y exactitud cuándo dos términos están relacionados. En el estándar, como norma general, se establece que dos términos están relacionados por asociatividad cuando uno de ellos puede formar parte de la definición del otro (figura 4.10).

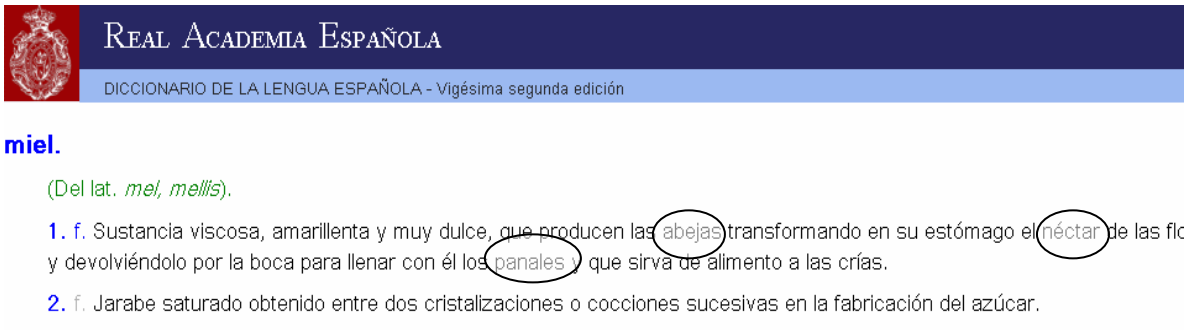


Figura 4.10. Aparición de los términos relacionados abeja, néctar y panales, en las definiciones del término miel

Además, los términos relacionados por asociatividad pueden no pertenecer a la misma jerarquía (figura 4.11).

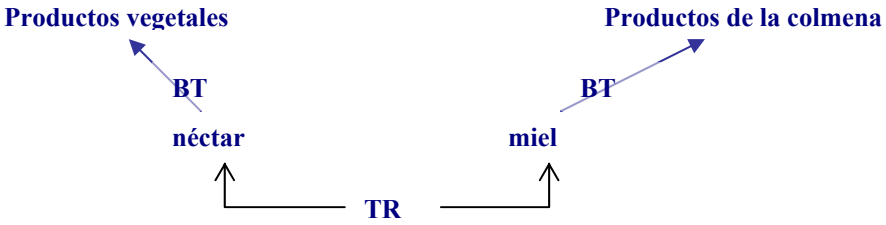


Figura 4.11. Relaciones de asociatividad entre términos pertenecientes a jerarquía diferentes

[26] http://es.wikipedia.org/wiki/AGROVOC

Además de los elementos básicos tradicionalmente definidos en un tesauro, términos, categorías y relaciones, en última versión del estándar ANSI/NISO Z39.19 (2005) se incorporan, para los tesauros de explotación, dos elementos más: los objetos de contenidos, tanto primarios como secundarios, y los índices.

4.3.1.4. Objetos de contenido

Los objetos de contenido corresponden a todo aquello que sirve de materia o asunto al ejercicio de las facultades mentales y que puede ser incluido en un sistema de recuperación de información, sitio web o cualquier otra fuente de información. Ejemplos típicos son los documentos, artículos, libros, etc., las páginas Web, los recursos didácticos digitalizados. Se distinguen dos clases de objetos de contenido, primarios y secundarios (figura 4.12). Los objetos primarios son los materiales en sí mismos, y los objetos secundarios son los metadatos que definen a los objetos primarios.

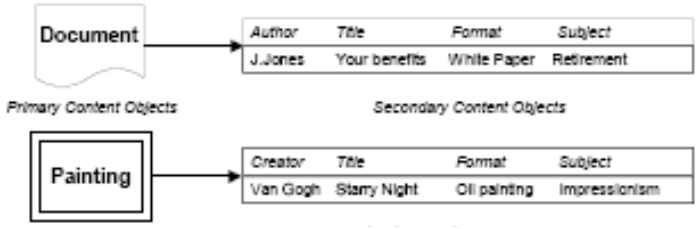


Figura 4.12. Objetos de contenido primario (izq.) y secundario (dcha)

4.3.1.5. Índices

Los índices son las estructuras que relacionan los términos del tesauro con los objetos de contenido (figura 4.13). La indexación es el proceso de escoger los términos del tesauro o descriptores que mejor definen a los objetos de contenido y asociarlos. Normalmente esta asociación se define y coloca en el objeto secundario, es decir, en los metadatos del objeto primario (Shapiro y Yan, 1996; Dalmau et.al., 2005). El proceso de indexación puede ser manual o automático, y puede utilizar términos procedentes de más de un vocabulario. La efectividad del proceso depende de que los vocabularios sigan el modelo estándar y los principios de diseño, significados únicos y bien definidos para los términos, términos sinónimos agrupados respecto de un término preferido, términos adecuados porque son los utilizados por los usuarios en las búsquedas.

Figura 4.13. Términos del tesauro ARIADNE para la indexación (campo Semantics) de un Recurso Educativo[27].

4.3.2. Acceso al contenido

Los estándares definen cinco modos de acceso al contenido de los tesauros:

1. presentación alfabética: el tesauro se presenta como una lista de términos ordenada alfabéticamente respecto de los términos, incluyendo los no preferidos (figuras 4.3 y 4.5). Desde cada término se accede directamente a su microestructura[28].
2. índice permutado: se trata de una variante de la presentación alfabética[29] que muestra una lista con todas las palabras individuales que componen los términos, simples o multipalabra, del tesauro, y un acceso a los términos que incluyen esa palabra (figura 4.14).

[27] El sistema de recuperación SILO (Search & Index Learning Objects) del repositorio ARIADNE utiliza estos términos del tesauro para indexar y buscar los objetos (además de otros puntos de acceso definidos en los metadatos como el título, autor, etc).
[28] Bien porque el término es un enlace a la microestructura o bien porque se incluye directamente a continuación del descriptor del término.
[29] A su vez se distingue entre índice permutado KWIC (KeyWord In Context) y KWOC (KeyWord Out of Context).

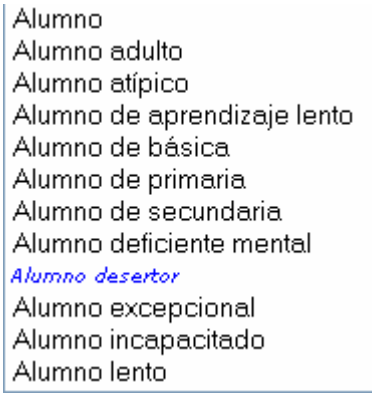


Figura 4.14. Índice permutado (término de búsqueda "alumno"). Tesauro SPINES de la UNESCO[30]

3. presentación jerárquica: complementa la alfabética, que sólo muestra un nivel jerárquico BT/NT para cada término, ampliando la vista al resto de los niveles de las jerarquías a las que pertenece el término (Aitchison et al., 2000; ANSI/NISO Z39.19, 2005) (figura 4.15).

Figura 4.15. Presentación de la jerarquía TE del término aprendizaje[31]

[30] http://spines.r020.com.ar/index.php
[31] Fuente: Tesauro Europeo de la educación, versión española, disponible en: http://www.freethesaurus.info/redined/es/index.php

4. presentación sistemática: visualiza la estructura global del tesauro. Se describe con detalle en la siguiente sección (figura 4.16)

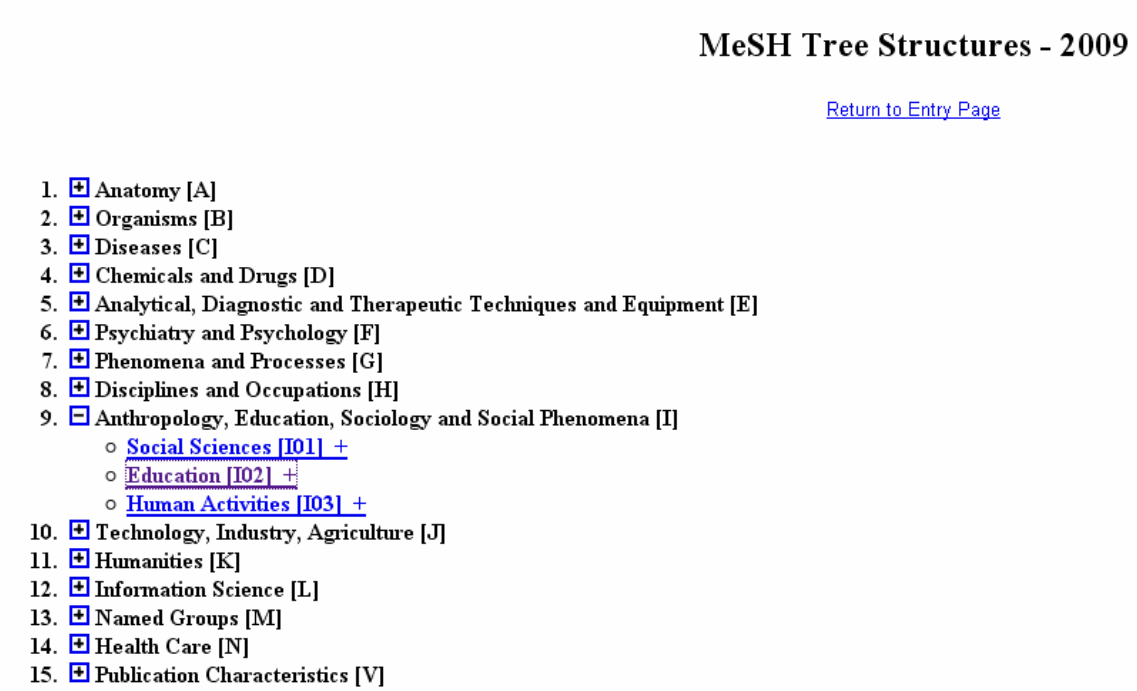


Figura 4.16. Presentación sistemática (facetada-jerárquica). MeSH Thesaurus[32]

5. Presentación Gráfica: presenta, en dos dimensiones, la macroestructura y microestructura del tesauro. Constituye un mapa terminológico y conceptual del alcance del tesauro. Debido a la limitación de espacio, normalmente sólo se muestra una parte de la macroestructura, seleccionada a partir de un término o categoría. Esta presentación permite la navegación para explorar, buscar y seleccionar los términos más adecuados a una consulta (figuras 4.17, 4.18 y 4.19).

[32] Fuente: Medical Subject Headings thesaurus, disponible en: http://www.nlm.nih.gov/mesh/2009/mesh_browser/MeSHtree.I.html

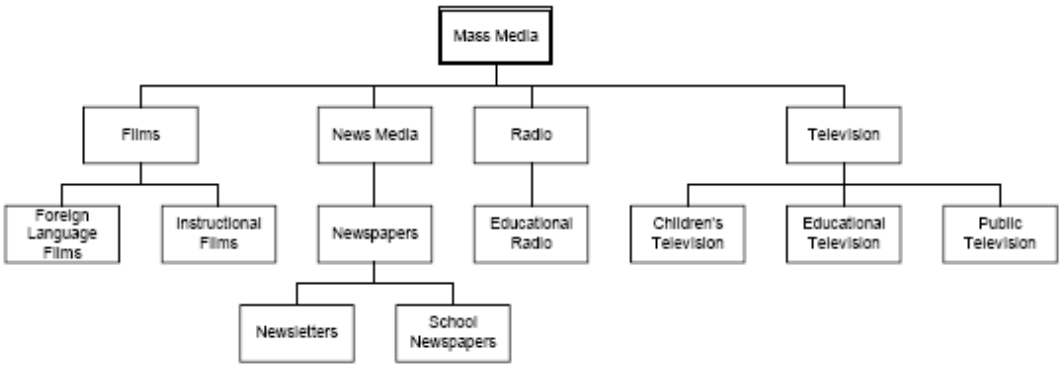


Figura 4.17. Vista parcial de la estructura jerárquica de hiponimia-hiperonimia del tesauro ERIC[33]

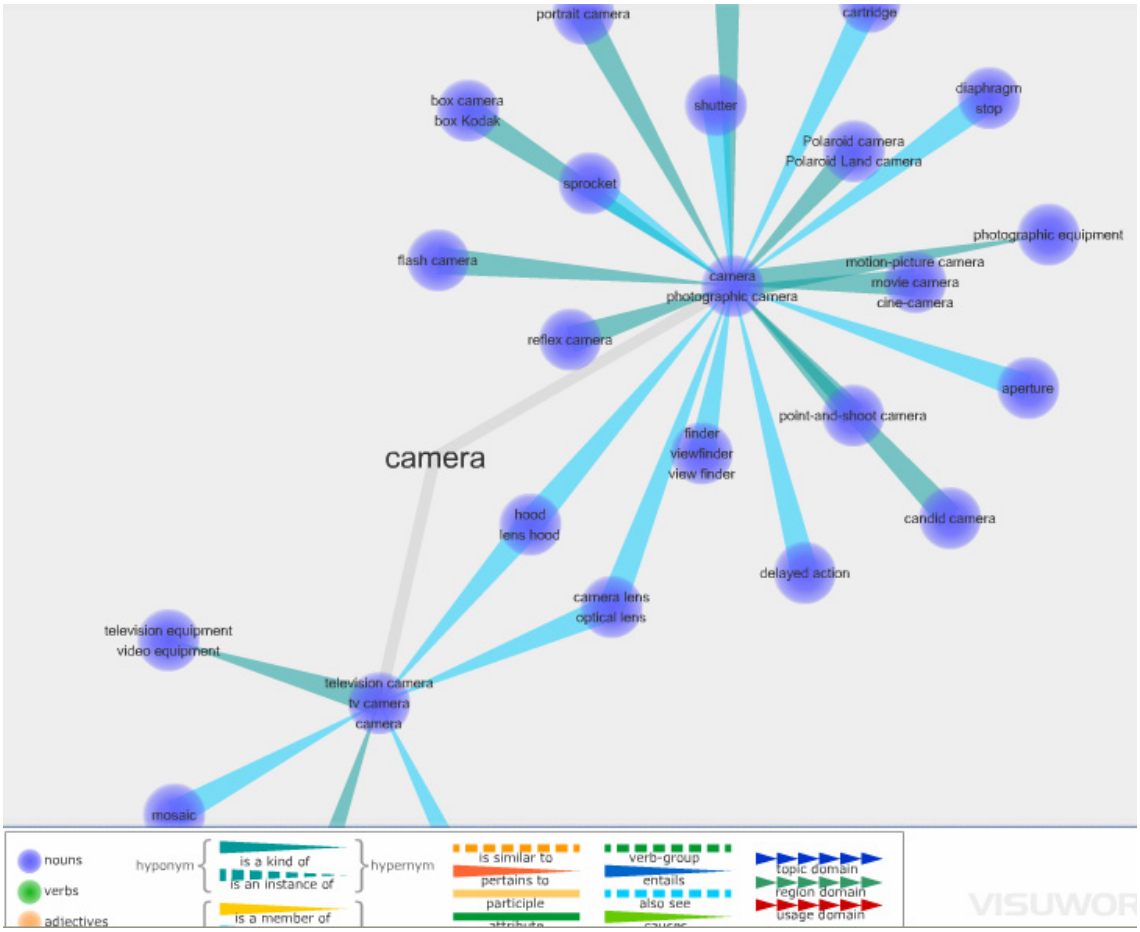


Figura *4.18. Presentación gráfica del vocabulario WordNet*[34]

[33] Fuente ANSI/NISO Z39.19, (2005).

[34] WordNet es un vocabulario estructurado como una red semántica. No la consideramos un tesauro puesto que no es un vocabulario controlado: múltiples conceptos para cada término y relaciones semánticas más ricas que las de los tesauros. Disponible en: http://www.visuwords.com/

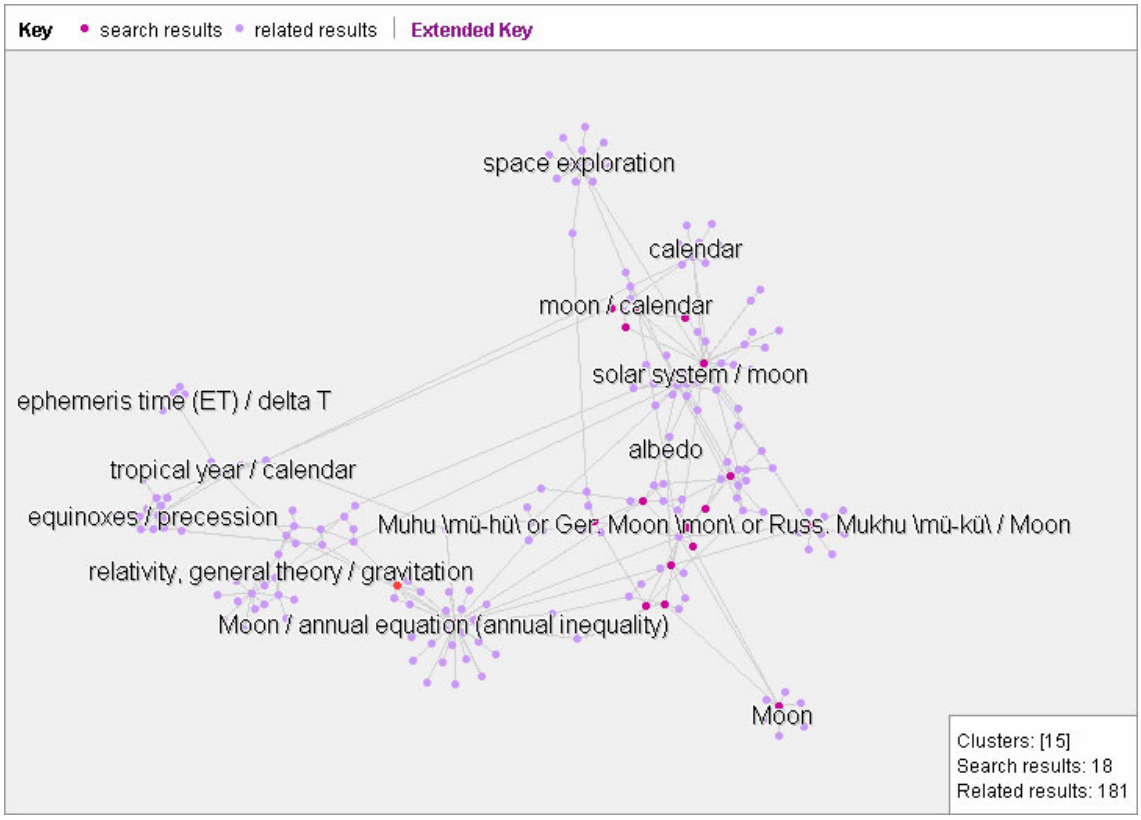


Figura 4.19. El tesauro como mapa terminológico-conceptual[35].

4.3.3. Operaciones de modificación

Los tesauros son un reflejo del lenguaje y, por lo tanto, son instrumentos dinámicos en permanente cambio (ANSI/NISO Z39.19, 2005). La necesidad de hacer modificaciones crece con el número de objetos de contenido que hay que indexar. El tesauro crece rápidamente en la fase inicial de construcción, hasta que se estabiliza. A partir de entonces las modificaciones son menos numerosas y el crecimiento es lento, excepto cuando el tesauro incorpora una nueva área temática.

Las modificaciones deben hacerse metódicamente para evitar inconsistencias que generan resultados confusos en la indexación y búsqueda en el tesauro (Aitchinson et al., 2000). Para cualquier modificación es recomendable hacer un análisis de la base de datos de objetos de contenido indexados por el tesauro para medir la repercusión de los cambios.

Los cambios pueden clasificarse, según su complejidad ascendente, en seis tipos (Aitchison et al., 2000):

[35] Los términos están visualmente agrupados en las categorías (clusters) del tesauro. Fuente ANSI/NISO Z39.19, (2005), example 161 p.86.

1) Actualizar la forma ortográfica de un término. Se trata de detectar algún error ortográfico en un término; esta modificación no afecta a otros términos o relaciones, pero si el término es redundante, es imprescindible asegurarse de que la actualización se realiza en todos los lugares donde aparece, para no perder la consistencia. Además, si el término es erróneo y se ha utilizado para indexar, debe: (i) borrarse el término erróneo (almacenándolo en un histórico), (ii) insertarse el término correcto, y (iii), enlazar el término correcto con el erróneo con una relación de tipo USE/USE PARA.
2) Actualizar el estatus de un término. Consiste en cambiar un término no preferido a preferido. En este caso, el término que figuraba como preferido en su clase de equivalencia debe "degradarse" a no preferido.
3) Borrar un término. Si un término no ha sido utilizado o es redundante se puede decidir borrarlo[36]. Cuando el borrado afecta a las microestructuras de las que formaba parte o se ha utilizado como término de indexación, el borrado no elimina el término, sino que lo marca como borrado y lo almacena en una estructura aparte denominada 'histórico' (figura 4.20). Los términos del 'histórico' están conectados con el tesauro bien mediante la relación USE con algún sinónimo o bien con la relación TE con algún hiperónimo. Finalmente, el sinónimo o el hiperónimo reemplazan al término borrado en la macroestructura.

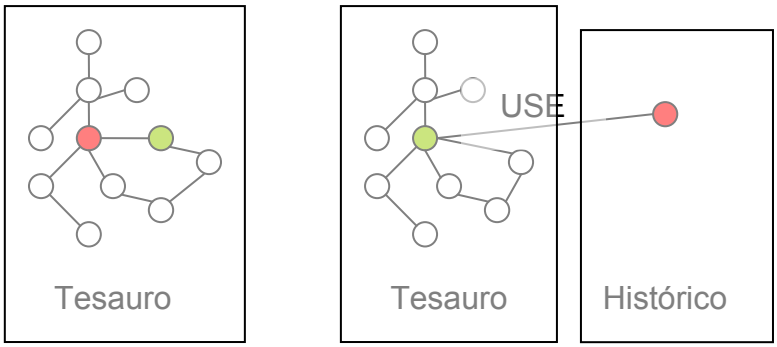


Figura 4.20. Borrado de un término (en color rojo) manteniéndolo en el histórico

4) Insertar o borrar relaciones. En el caso de que las relaciones sean erróneas se eliminan y se inserta una nueva relación. Si son relaciones no útiles simplemente se eliminan.

[36] Estas decisiones suelen estar basadas en estadísticas de uso de los términos en la indexación y en las búsquedas.

5) Insertar nuevos términos. En el caso de que durante el indexado y/o búsqueda, los usuarios utilicen términos externos al tesauro puede considerarse la posibilidad de su inclusión. Normalmente se hace una inclusión temporal, término candidato, y, después de un tiempo de prueba, si el término es útil se confirma como definitivo. Para mantener esta información se añaden campos a la entrada o microestructura de un término con las fechas de inclusión. Si es candidato, esta modificación implica: (i) preparar convenientemente la microestructura teniendo en cuenta la macroestructura, y (ii) conectar la nueva microestructura en el lugar conveniente de la macroestructura. Y
6) Actualizar la macroestructura. Es la modificación más complicada si el tesauro está ya construido, porque no es fácil garantizar la consistencia de la información. Resulta de gran ayuda, para localizar los elementos afectados, disponer de los esquemas de datos conceptuales y de implementación de datos que se hayan diseñado.

En resumen, los estándares de construcción de tesauros se limitan a proponer un sistema de descripción, visualización y actualización de los vocabularios controlados que incluye las relaciones semánticas básicas, las normas para la selección de las unidades léxicas que forman los descriptores o términos preferidos, y las formas de presentación del contenido. Este sistema tiene como objetivos: (i) el control de la sinonimia y la polisemia para que el tesauro sea un sistema eficaz de indexación y búsqueda de objetos de contenido, y (ii) el establecimiento de una nomenclatura común que permita el intercambio y reutilización de tesauros. Debe constituir, por lo tanto, el referente para los modelos de datos informáticos que se utilicen para la construcción de tesauros monolingües.

4.4. La aplicación de los modelos alfabético y sistemático de los estándares a la construcción de los tesauros

La aproximación más tradicional de construir tesauros está basada en los modos de presentación alfabética y sistemática para facilitar el acceso y garantizar la coherencia del contenido de los tesauros, tanto impresos como digitales. La razón es que la forma de presentación del contenido determina la forma de acceso al mismo. En esta aproximación, la estructura y la presentación del tesauro es una misma cosa: la

presentación del tesauro es exactamente su estructura y su estructura es la forma de presentar el tesauro. En esta sección vamos a profundizar un poco más en estos modelos de tesauro.

4.4.1. El modelo alfabético

La lista alfabética es, probablemente, la forma más fácil de construir y reproducir un tesauro. Se estableció, por primera vez, en 1967 con el Tesauro de Términos Científicos y de Ingeniería (TEST)[37]. El esquema de datos es, simplemente, una lista ordenada alfabéticamente (figura 4.21) de "fichas de términos", descriptores o no, donde cada ficha tiene la siguiente estructura:

TÉRMINO PREFERIDO (DESCRIPTOR)
NA (SN) notas de ámbito o defínción
UP (UF) referencias a los términos equivalentes (términos no preferidos)
TG (BT) referencias a los términos más generales (hiperónimos)
TE (NT) referencias a los términos más específicos (hipónimos)
TR (RT) referencias a los términos relacionados

TÉRMINO NO PREFERIDO (NO DESCRITPTOR)
USE TÉRMINO PREFERIDO (DESCRIPTOR) (y TÉRMINO PREFERIDO 2)

[37] Engineers Joint Council. (1967). *Thesaurus of engineering and scientific terms;* es una lista de terminos científicos y de las Ingenierías cuyo objetivo es servir de referencia para el indexado y recuperación de información técnica.

Figura 4.21. Esquema de datos alfabético del Tesauro de Redes de Ordenadores (Martínez y García, 2006)

Las jerarquías generadas por las relaciones TG y TE están implícitas en este esquema y no es posible examinar, a partir de un término, todos los términos genéricos y específicos de la jerarquía o jerarquías a las que pertenece. Por esta razón, en algunos tesauros alfabéticos se añade una información que, aunque es redundante, ayuda al lector a reconstruir las jerarquías a las que pertenece el término: (i) el término inicial, o raíz, de la jerarquía llamado Cabeza de Serie (CS) o en inglés Top Term (TT), y (ii) dos o más niveles de términos genéricos (TG) o de términos específicos (TE). En cualquier caso, es información redundante, con el consiguiente peligro de posibles inconsistencias en caso de modificaciones[38].

Este modelo es apropiado para tesauros en soporte papel, impreso o digital, pero no es adecuado para su utilización en la explotación de contenidos, porque muestra la información fragmentada. El usuario es el que tiene que reconstruir el "mapa global" de

[38] Supongamos, por ejemplo, que se decide cambiar un término en un determinado punto de la jerarquía. Esto supondría localizar todas las entradas que están haciendo referencia a ese término y cambiarlas. Un error en alguna generaría inconsistencias difíciles de detectar.

términos y relaciones o macroestructura del tesauro. En definitiva, no facilita al lector la comprensión del dominio.

4.4.2. El modelo sistemático

La representación sistemática, también llamada de clasificación o temática, pretende una organización de los términos en categorías y/o en jerarquías de acuerdo a sus significados e interrelaciones lógicas (figura 4.22). El esquema sistemático muestra la estructura global o macroestructura del tesauro (ISO 2788, 1986), haciendo explícitas las relaciones entre las jerarquías, categorías y grupos de términos.

Figura 4.22 Presentación sistemática. Vista parcial de la jerarquía de la faceta ARQUITECTURA Y DISEÑO DE REDES del Tesauro de Redes de Ordenadores (Martínez y García, 2006)

El diseño sistemático genera un esquema de tesauro con tres partes diferenciadas (UNE 50106, 1990) (figura 4.23):

1) la sección sistemática: formada por las categorías o jerarquías globales de los términos organizados según sus significados e interrelaciones lógicas. Es el esquema conceptual que describe la materia o dominio de conocimiento utilizando ciertos términos preferidos del lenguaje natural (figuras 4.24 y 4.25);
2) la sección alfabética o índice alfabético, que guía al usuario a la parte o partes apropiadas en la sección sistemática. Incluye todos los términos del lenguaje natural, preferidos y no preferidos, que se refieren al dominio; y
3) el sistema de referencias, conecta las dos partes anteriores. Normalmente es un sistema de códigos (numéricos o alfabéticos) que se asignan a cada término preferente para que sirvan de reenvío desde la parte alfabética (ANSI/NISO Z39.19, 2005; Lancaster, 1986). En el caso de tesauros digitales e hipertextuales el sistema de referencia se construye mediante un sistema de enlaces hipertextuales.

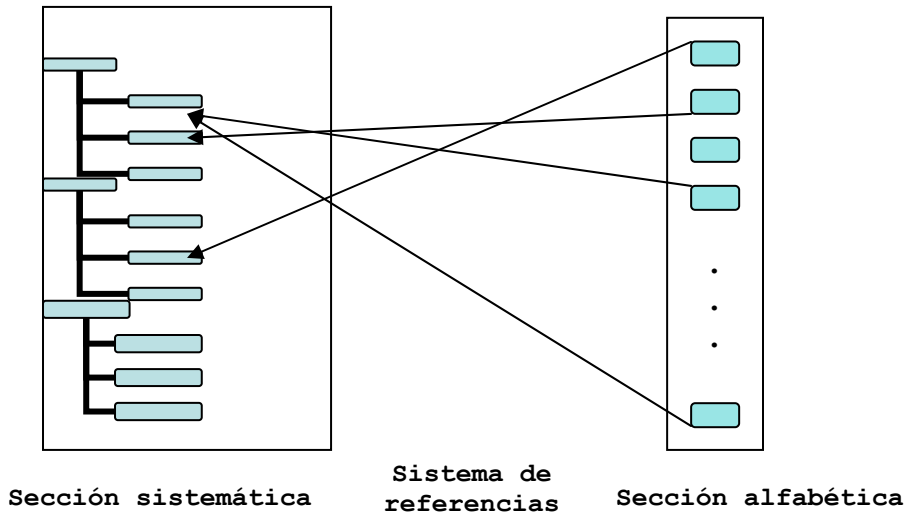


Figura 4.23. Las tres partes del esquema sistemático del tesauro: sección sistemática (izq.), sistema de referencias (flechas) y sección alfabética (dcha.)

La sección sistemática se diseña utilizando categorías, jerarquías y listas, y entradas. Las *categorías*, que constituyen la división inicial y principal del tesauro, se obtienen mediante técnicas de clasificación. Se distinguen tres tipos de categorías:

a) por campos o disciplinas. El tesauro se organiza en campos temáticos que describen conceptualmente el ámbito del tesauro. Para ello, es necesario definir una categoría por cada campo temático y las relaciones de subordinación que mantienen entre ellos. La estructura resultante es una jerarquía (figura 4.24), o

un conjunto de jerarquías, polijerarquía, que pueden solaparse por compartir algunos términos.

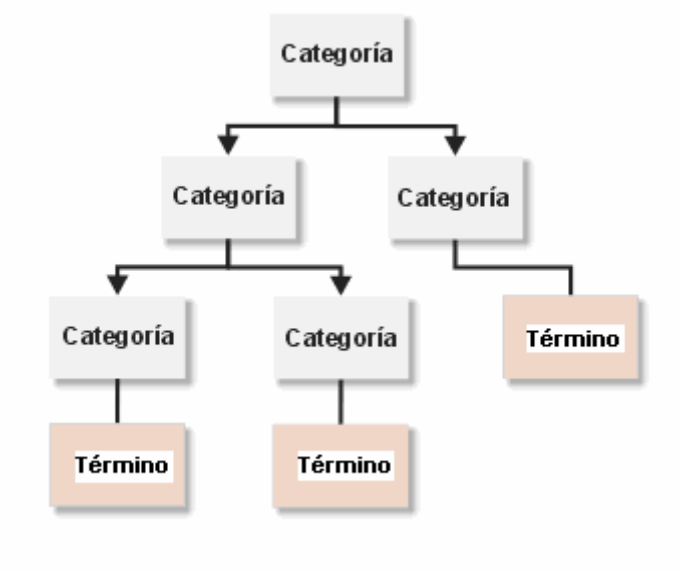


Figura 4.24. Esquema de una clasificación por campos[39]

Este tipo de clasificación permite agrupar los términos conceptualmente relacionados y ofrecer al usuario un esquema de organización del dominio de conocimiento del tesauro que puede ser de gran ayuda en la búsqueda. Sin embargo, tiene como inconveniente la subjetividad inherente a cualquier esquema conceptual de organización del conocimiento, lo que puede generar problemas como: (i) esquemas de organización que no se corresponden con las estructuras mentales del usuario; y (ii) términos asignados a categorías diferentes en distintos tesauros, o dificultad en reconciliar esquemas temáticos de distintos tesauros que dificultan el intercambio y compartición de colecciones de objetos de contenidos pertenecientes a un mismo dominio.

Ejemplos conocidos son el tesauro ERIC de recursos educativos (figura 4.26), tesauro SPINES de la UNESCO y las clasificaciones bibliográficas LCC, LCSH, DDC y UDC[40].

b) por facetas. Los términos se dividen en categorías disjuntas denominadas facetas utilizando sólo una característica (o principio) de división cada vez[41], (figura 4.25)

[39] Modificado de http://www.nosolousabilidad.com/articulos/clas_facetadas1.htm
[40] LCC (Library of Congress Classification) http://lcweb.loc.gov
LCSH (Library of Congress Subject Headings)http://lcweb.loc.gov
DDC (Dewey Decimal Classification)http://www.oclc.org/oclc/fp/
UDC (Universal Decimal Classification)http://zeus.slais.ucl.ac.uk/udc/

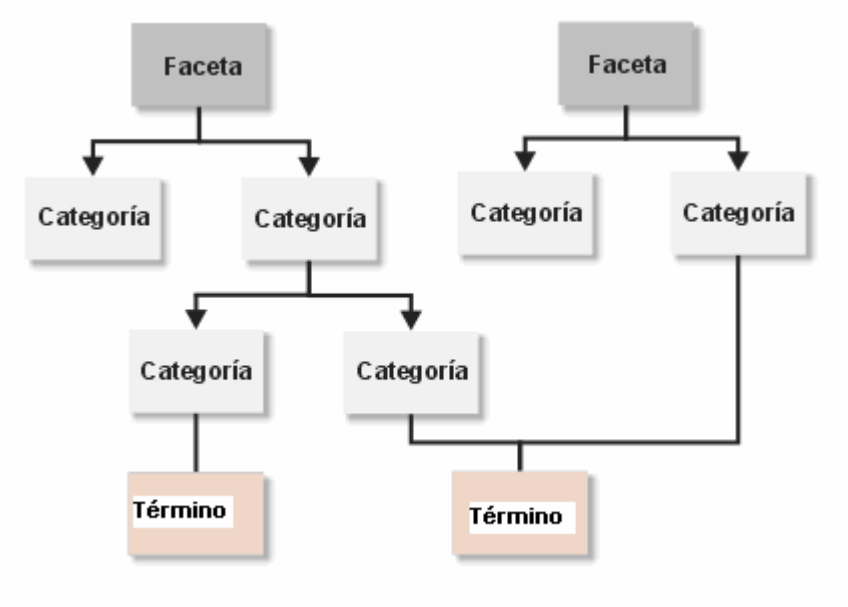


Figura 4.25. Esquema de una clasificación facetada[42]

Por ejemplo, la clasificación mostrada en la figura 4.6 divide los hipónimos del término inglés *painting* del *Art and Architecture Thesaurus* en tres facetas o categorías disjuntas: *by form*, *by location or context* y *by material o technique*. Entre otros ejemplos en lengua española contamos con el Tesauro de Psicología del CINDOC y el Tesauro de Educación Superior de la UCM.

c) enfoque combinado. En la práctica es habitual combinar los dos tipos anteriores. Así, un tesauro organizado primariamente en campos temáticos puede subdividirse en facetas (o viceversa)[43].

En cualquiera de los casos se asegura que los términos equivalentes aparezcan juntos dentro de la misma categoría.

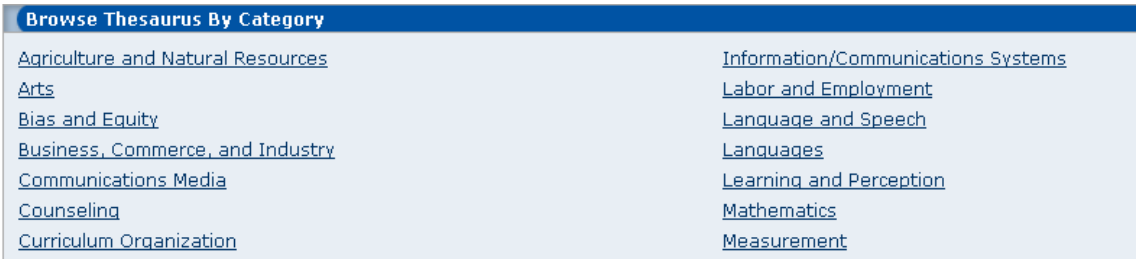


Figura 4.26. Categorías (por campos) principales del tesauro ERIC[44]

[41] Fue el documentalista S.R. Ranaganathan quien desarrolló y popularizó este tipo de clasificación y lo aplicó en su clasificación Colon (Ranaganathan, 1987).
[42] Modificado de http://www.nosolousabilidad.com/articulos/clas_facetadas1.htm
[43] Ejemplo de UNE 50106, (1990): el campo "Educación" puede dividirse en facetas que representen categorías básicas como: operaciones (por ejemplo "enseñanza"), procesos (por ejemplo "aprendizaje") y agentes (por ejemplo "profesores").

Las jerarquías y listas son las estructuras que organizan los términos de cada una de las categorías. Las jerarquías, o árboles, estructuran, en primer lugar, los términos hipónimos e hiperónimos y las listas, en segundo lugar, los términos relacionados por asociatividad (4.27).

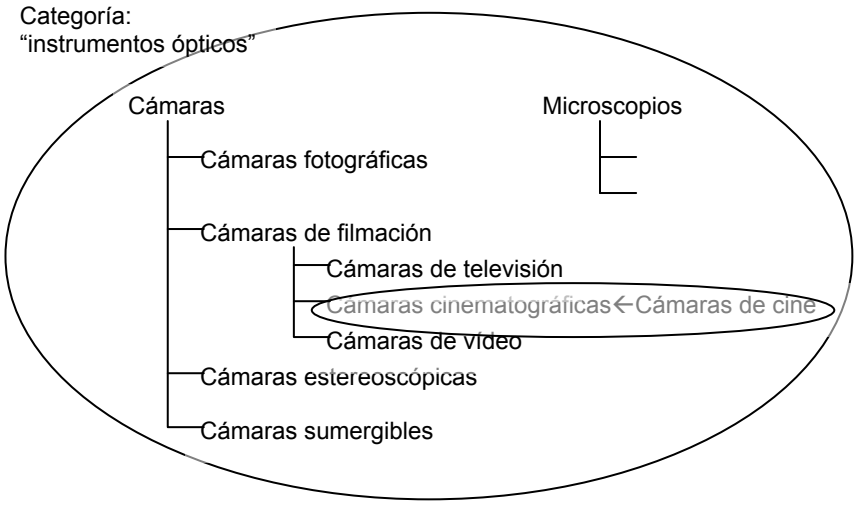


Figura 4.27. Esquema estructural de una categoría[45] [46]

Una jerarquía o árbol es una estructura de datos formada por un elemento raíz y cero o más subjerarquías. Las subjerarquías, a su vez, están ordenadas (DADS, 2007). Los términos dentro de las categorías se organizan teniendo en cuenta las relaciones de generalidad y especificidad que mantienen entre ellos. Las jerarquías se ordenan por orden alfabético respecto de sus términos raíces. Se utiliza, como hemos visto, algún tipo de marca para señalar estas relaciones: sangrado de hipónimos o las etiquetas estándares BT y NT (figura 4.15 y 4.16).

Una lista o array de términos es un conjunto de términos organizados secuencialmente y accesibles por su posición. Dentro de una categoría y para un determinado término, situado en su lugar correspondiente de la o las jerarquías de hiponimia-hiperonimia, los términos relacionados con él por asociatividad se organizan en listas, normalmente, por orden alfabético (figura 4.28).

[44] Disponible en: http://www.eric.ed.gov/thesaurus
[45] Creada a partir de ejemplos de (Lancaster, 1986) y (UNE 50106, 1990)
[46] Esta categoría, "instrumentos ópticos" se estructura mediante dos jerarquías ordenadas alfabéticamente: "Cámaras" y "Microscopios". La jerarquía "Cámaras" tiene, además, un conjunto de términos equivalentes: "Cámaras cinematográficas", que es preferido, y "Cámaras de cine", no preferido.

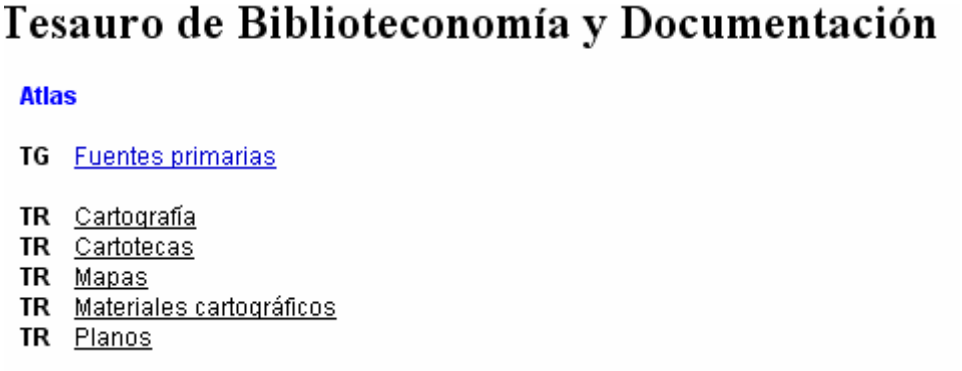


Figura 4.28. Listas de términos relacionados (TR)[47]

Finalmente, el último elemento de la sección sistemática son *las entradas*. Los términos, además de su descriptor, pueden tener información no relacional propia: notas de ámbito, fechas de propuesta e inclusión, personas que proponen el término, que aprueban la inclusión, etc. Esta información se estructura en la entrada, que es como una ficha con un conjunto de pares atributo y valor. La mayor parte de la información de la entrada no es visible para los usuarios. Además, conviene distinguir entre la entrada de la sección sistemática y la entrada de la sección alfabética. La información de la entrada alfabética incluye a la de la entrada sistemática, puesto que la entrada alfabética incluye toda la información relativa a un término: relacional y no relacional (figura 4.4).

La sección alfabética se construye a partir de la sistemática (Aitchison et al., 2000), como se ha descrito anteriormente en la sección anterior, 4.4.1.

El tesauro se visualiza reproduciendo su estructura: alfabética y sistemática. Las consultas, por lo tanto, están predeterminadas: desde la sección alfabética se accede a la microestructura del tesauro, desde la sistemática, a la macroestructura. Las modificaciones se realizan de la forma descrita en el estándar de construcción de tesauros (sección 4.3.3).

En resumen, este modelo utiliza como elementos estructurales para organizar los términos de un tesauro a: (i) las categorías, (ii) las jerarquías (relación hiponimia-hiperonimia, holonimia-meronimia) y listas (relaciones de asociatividad y equivalencia), y (iv) la entrada. Las categorías no forman parte del vocabulario en sí, son metainformaciones semánticas en este caso, y contienen términos, normalmente estructurados en jerarquías que están ordenadas alfabéticamente respecto del término raíz. Cada uno de los términos de la jerarquía puede contener una o dos subestructuras

[47] Disponible en Tesauros y Glosarios IEDCYT en línea: http://thes.cindoc.csic.es/

de tipo lista, con los términos que están relacionados semánticamente con él y otro con los términos equivalentes (sinónimos, variantes, etc). Los términos se clasifican en preferidos y no preferidos, siendo los preferidos los que sirven para la indexación, y los no preferidos los que apuntan a los preferidos, aunque sirven para ampliar las posibilidades de búsqueda. Cada término preferido puede tener, además, un conjunto de propiedades no relacionales como la definición (nota de ámbito), un código (sistema de referencias), etc.

Para concluir, destacamos, desde el punto de vista de la organización de la información y de objetos de contenido, dos cuestiones sobre el modelo sistemático:

1) los objetos de contenido están asociados a términos, por lo tanto, están localizados en una estructura conceptual, de categorías y jerarquías de términos, que representa el dominio del conocimiento y es la macroestructura del tesauro. Esto facilita la aplicación de diversas formas de localización de los objetos buscados: navegación, postcoordinación, precoordinación.

2) el modelo sistemático tiene la ventaja de que obliga a llevar a cabo un planteamiento de diseño global del tesauro, de carácter macroestructural, no sólo de relaciones terminológicas. Esta circunstancia, como hemos visto en la primera sección de este capítulo, (i) garantiza una mayor consistencia en el contenido del tesauro; (ii) facilita la interoperatividad, puesto que puede esperarse un mayor nivel de coincidencia entre diferentes centros o aplicaciones; y (iii) es más efectiva desde el punto de vista constructivo, de mantenimiento y de uso. Sin embargo, exige un esfuerzo mayor en el análisis y diseño de un esquema del tesauro para garantizar que los conceptos similares aparezcan juntos y en sus posiciones correctas respecto de unos criterios de organización lógica claramente establecidos y aplicados. Además, no se puede asegurar que un esquema de datos ajustado a un dominio sea, al mismo tiempo, suficientemente expresivo y flexible como para asimilar todos los cambios futuros del tesauro. En este sentido, *la experiencia demuestra lo difícil que es obtener esquemas sistemáticos sin problemas estructurales y durables toda la vida del tesauro* (Arano y Codina, 2004)[48]

[48] Los autores presentan un análisis de patologías estructurales (como la desnaturalización jerárquica o la clasificación cruzada) cuyo origen está en un diseño poco riguroso a nivel macroestructural. En el capítulo siguiente insistiremos en esta cuestión.

4.5. Resumen y conclusiones del capítulo

En este capítulo se han revisado los fundamentos teóricos sobre construcción de tesauros que se han ido elaborando de forma empírica durante décadas. Estos fundamentos tratan de (i) las características de los tesauros y sus requisitos, con especial énfasis en los tesauros de explotación, (ii) el modelo estándar de contenido, de acceso y de modificación de los tesauros de explotación, y (iii) los modelos de construcción tradicionales que son anteriores a la informatización de los tesauros.

Los tesauros se caracterizan por cuatro propiedades: 1) su naturaleza altamente relacional y en permanente cambio; 2) su estructura organizada en macroestructura y microestructura; 3) los elementos de contenido que son, fundamentalmente, los términos, las relaciones semánticas, las categorías y, en los tesauros de explotación, también los objetos de contenido y los índices; 4) los tipos de presentación del contenido que están establecidas en los estándares y son: alfabética, índice permutado, jerárquica, sistemática y gráfica. Los tesauros, además, deben cumplir tres condiciones 1) propias de su naturaleza, como son evitar la ambigüedad, controlar la sinonimia, representar con exactitud las relaciones y utilizar los términos más adecuados para representar los conceptos u objetos de un dominio; 2) condiciones relativas al dominio, como la cobertura, el alcance y el tipo de usuario; y 3) en el caso de los tesauros de explotación, las condiciones de expresividad, flexibilidad, efectividad y accesibilidad.

Los estándares de construcción de tesauros tienen como objetivo sistematizar el contenido, la presentación y las reglas de modificación de los tesauros. El contenido de un tesauro corresponde a términos organizados en categorías y en redes de relaciones semánticas. Las relaciones semánticas básicas son de (i) tipo jerárquico de término general-específico (TG-TE), (ii) tipo término equivalente (USE-USEPARA), y (iii) tipo término relacionado (TR). Los tesauros de explotación contienen, además, objetos de contenido asociados a los términos por medio de los índices.

Los modos de presentación definen, en realidad, los modos de acceso al tesauro más eficaces para que una persona encuentre el término u objeto de contenido deseado. La presentación más básica es la alfabética; pero, normalmente, disponen de más de un modo, por ejemplo, alfabético y sistemático, porque, al ser complementarios, mejora la eficacia del tesauro.

La definición de un conjunto de reglas para controlar las modificaciones, demuestra lo habitual que es en un tesauro las actualizaciones. Las reglas de modificación tratan de garantizar la consistencia del tesauro y facilitar las actualizaciones; sin embargo, éstas son difíciles y costosas, por lo cual no se realizan de forma continua, sino en periodos de tiempo de años. En definitiva, la sistematización estándar tiene como fin (i) garantizar que el tesauro sea un sistema eficaz de indexación y búsqueda de objetos de contenido, y (ii) establecer una nomenclatura común que permita el intercambio y reutilización del contenido de los tesauros y de los objetos de contenido. Debe constituir, por lo tanto, un referente para los modelos de datos informáticos que se utilicen en la construcción de los tesauros de explotación.

Los dos modelos tradicionales de construcción de tesauros, de los que se han ido derivando los modelos estándares, son el modelo alfabético y el modelo sistemático. El modelo alfabético es el más antiguo, el más sencillo de aplicar y el más barato, pero tiene dos inconvenientes importantes: (i) muestra el contenido del tesauro fragmentado, y (ii) es difícil de mantener la consistencia del contenido porque el esquema global está implícito y porque la información de las relaciones tiene que repetirse en las entradas de cada uno de los términos que participan en la relación. Es un modelo adecuado para tesauros con macroestructuras sencillas, que tengan pocas actualizaciones, y que se van a presentar en soporte papel impreso o digital.

El modelo sistemático es más elaborado, exige un esfuerzo mayor de diseño e implementación que el alfabético para obtener y construir el esquema sistemático del ámbito del tesauro, y no es fácil de actualizar cuando las modificaciones afectan al esquema. Sin embargo presenta tres ventajas: (i) aporta un esquema terminológico-conceptual del ámbito del tesauro que facilita la comprensión del dominio y la búsqueda aplicando múltiples estrategias basadas en la postcoordinación o precoordinación de términos y en la navegación; (ii) facilita el control de la consistencia del contenido del tesauro, tanto en la construcción como en el mantenimiento del mismo; y (iii) facilita la interoperatividad, porque el esquema del tesauro permite interpretar e intercambiar el contenido entre diferentes centros o aplicaciones. En definitiva, es un modelo adecuado para los tesauros de explotación siempre que se construya, con un modelo de datos informático, el esquema sistemático y que no se vayan a realizar actualizaciones que afecten a este esquema.

Los métodos de construcción de tesauros, tratados en el capítulo 7, se basan en este modelo para construir, de forma inductiva o deductiva, el esquema sistemático con el

que se organiza el contenido del tesauro. Después, se aplican modelos informáticos para reproducir este diseño y construir los tesauros de explotación. Los modelos informáticos se tratan en el siguiente capítulo.

Capítulo 5

Los modelos informáticos para la construcción de tesauros de explotación

En Informática los modelos de datos proporcionan un mecanismo para describir, estructurar y procesar, de forma eficiente, grandes cantidades de datos interrelacionados. Actualmente, el contenido de los tesauros de explotación se construye con modelos de datos informáticos combinados con el modelo estándar de tesauro. De esta forma se asegura la representación, el acceso y el procesamiento automático de su contenido y la integración en otros sistemas informáticos. En este capítulo se revisan los modelos de datos más utilizados para la construcción de tesauros, y se analizan a partir de los requisitos específicos de los tesauros de explotación, recogidos en el capítulo anterior: 1) si son suficientemente expresivos como para representar el contenido de cualquier tesauro, 2) si son suficientemente flexibles como para garantizar la introducción continua de modificaciones que pueden afectar no sólo al contenido sino también a la estructura, y 3) si son eficaces para visualizar y acceder al contenido del tesauro.

Los modelos que revisamos se han seleccionado porque son los más utilizados o porque son estándares o propuestas de estándares. También se ha hecho hincapié en los modelos que provienen de la tecnología educativa, por ser cercanos al área de aplicación de este trabajo de investigación.

5.1. La informatización de los tesauros

En los modelos tradicionales los tesauros se concibieron como un conjunto de términos organizados semánticamente utilizando, principalmente, categorías y/o relaciones semánticas. Los términos representan un único concepto, puesto que no debe existir ambigüedad ni polisemia. Esta concepción, sin embargo, ha cambiado con la informatización de los tesauros y con sus nuevas aplicaciones como sistemas lingüísticos de representación del conocimiento común, compartido, necesario para la interpretación y descripción automática de información, recursos y sistemas.

En este nuevo escenario informático existen diferentes formas de concebir un tesauro, que van desde los diseños terminológicos hasta los ontológicos. En el primer caso, el tesauro es una estructura de términos relacionados, mientras que, en el segundo caso,

son estructuras de conceptos relacionados que pueden estar asociados a términos. En este sentido, existe una frontera, a veces difusa, entre los tesauros y las ontologías que hace difícil la distinción. Utilizando como referencia la clasificación propuesta para los tesauros en formato XML, de Matthews et al. (2003), se puede distinguir entre dos tipos de tesauros, (i) los basados en términos, y (ii) los basados en conceptos (figura 5.1). Cuando el tesauro está basado en términos, contiene términos y relaciones, que pueden estar organizados en categorías y relaciones en las aproximaciones terminológicas, o en clases y subclases en las aproximaciones terminológica-conceptuales. Cuando el tesauro está basado en conceptos se construye utilizando conceptos y relaciones, y se conoce como aproximación ontológica; o en conceptos, términos para denominar a los conceptos y relaciones, y se llama aproximación ontológica-terminológica, tal y como se muestra en el esquema de la figura 5.1:

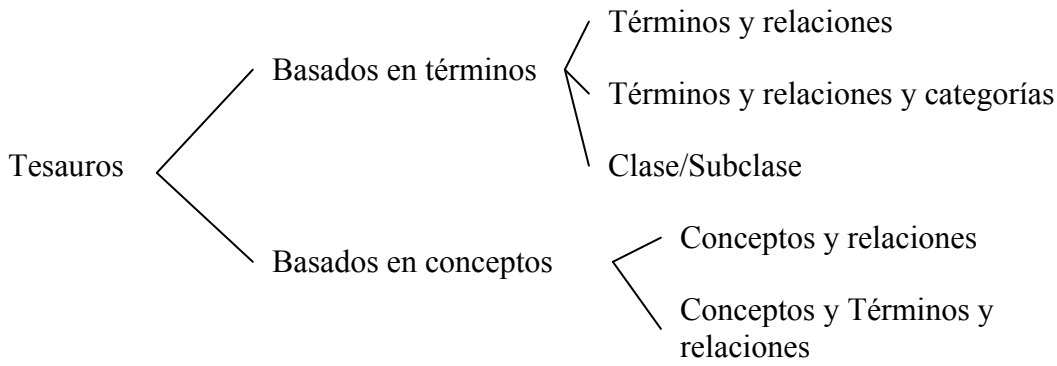


Figura 5.1. Clasificación de modelos de tesauros (adaptada de Matthews et al., 2003).

En este capítulo se presentarán varios ejemplos de tesauros, algunos están basados en términos y otros en conceptos. Por ejemplo, el tesauro CERES[1] (figura 5.21) está basado en términos y se ha construido conforme a los modelos estándares de tesauros, con términos y relaciones semánticas TG/TE, TR y USE. El tesauro Agrícola de la NAL[2] (figuras 5.19) y el tesauro ELSST[3] (figuras 5.23 y 5.24) están basados en conceptos. Se han construido con jerarquías IS-A de conceptos con atributos y relaciones semánticas TG/TE y TR. Utilizan una aproximación basada en ontologías

[1] Disponible en: http://ceres.ca.gov/thesaurus/Overview.html
[2] Tesauro de la National Agricultural Library. Disponible en: http://agclass.nal.usda.gov/dne/search.shtml
[3] European Language Social Science Thesaurus (Balkan et al., 2002).

terminológicas, que están a mitad de camino entre los tesauros y las ontologías[4]. El objetivo de esta concepción ontológica de los tesauros es disponer de un esquema conceptual común que permita la interoperabilidad entre sistemas que utilizan tesauros diferentes.

Para que los diferentes tipos de tesauro sean compatibles es importante utilizar los modelos estándar de construcción de tesauros con los modelos de datos informáticos. Los modelos estándar aglutinan un conjunto de elementos de contenido, formas de acceso y reglas de modificación que deben ser representadas y manipuladas con los modelos informáticos (ver capítulo 4). Asimismo, al igual que los modelos estándar distinguen entre la macroestructura y microestructura[5], los modelos de datos informáticos también deben distinguir estos los dos niveles estructurales[6] (Aitchison et. al, 2000; Gibbon, 2000). La microestructura contiene la información relacional y no relacional de cada término (figura 5.2), mientras que la macroestructura describe cómo se relacionan todas las microestructuras, cada término y sus interrelaciones, para formar un esquema global del tesauro que reproduce la "forma" del dominio de conocimiento del tesauro (figura 5.3).

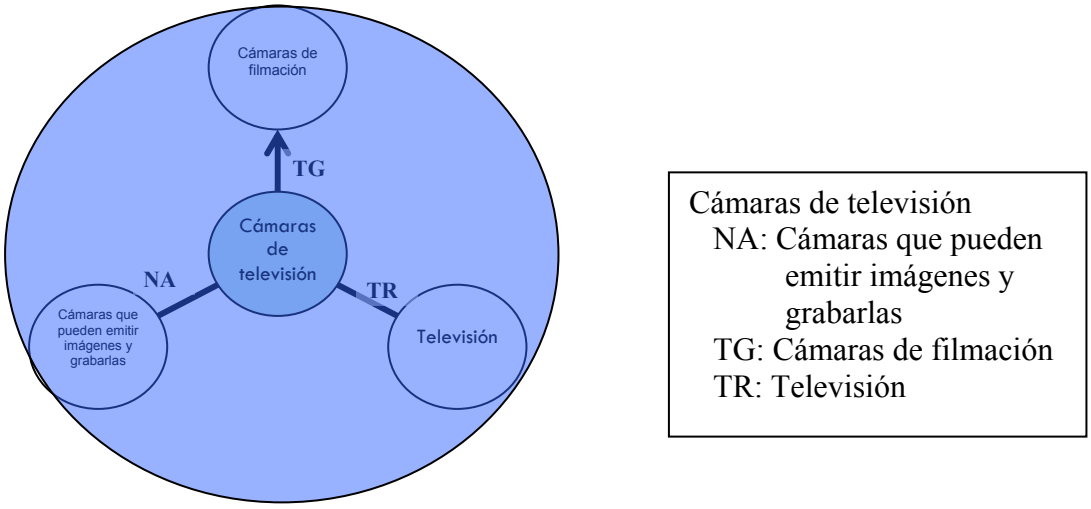


Figura 5.2. Microestructura del elemento "Cámaras de televisión"

[4] Algunos autores cuestionan que puedan llamarse ontologías por la imprecisión de las relaciones. Por ejemplo, la relación TG/TE sirve para denotar varios tipos –hiperonimia-hiponimia, holonima, instancia e implicación- con significados diferentes (Matthews et al., 2003).

[5] El modelo alfabético hace hincapié en la microestructura y el modelo sistemático prioriza la representación de la macroestructura, a partir de la cual también es posible acceder a la microestructura de los términos.

[6] Sin embargo, los tesauros tradicionales tienen un inconveniente que debe evitarse en los tesauros informatizados: no existe una independencia real entre ambos componentes: la microestructura aparece embebida en la macroestructura, modelo sistemático, o bien la macroestructura no aparece definida explícitamente, modelo alfabético. Esto dificulta el mantenimiento de estos tesauros, porque los cambios en la microestructura pueden afectar de forma no previsible a la macroestructura y viceversa.

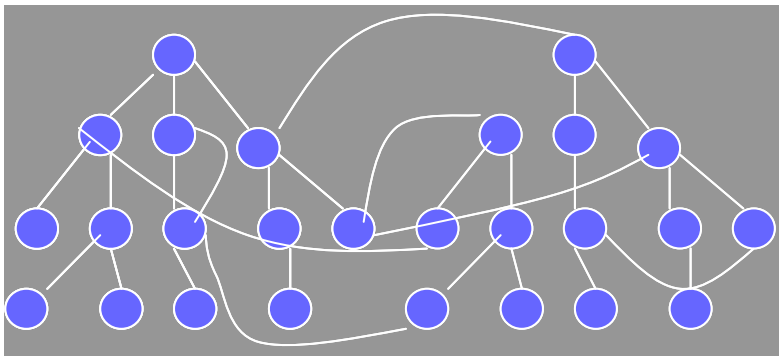


Figura 5.3. Composición de las microestructuras, que se representan por bolas azules, para formar la macroestructura.

Los modelos de datos informáticos proporcionan, como mínimo, un conjunto de conceptos para ayudar a definir el esquema de datos del tesauro. Los modelos de datos más completos también permiten definir:

1) un conjunto de reglas para la inserción, actualización y borrado de información que aseguren el mantenimiento de la integridad;
2) un conjunto de operaciones para consultar y manipular el contenido.

Este conjunto de operaciones básicas necesarias para acceder al contenido de los tesauros es:

i) obtener la microestructura de un término dado;
ii) obtener la jerarquía o jerarquías a las que pertenece un término dado;
iii) obtener todas las redes de términos generadas por una relación semántica dada;
iv) obtener todos los términos relacionados con una determinada relación semántica y con un término dado. Por ejemplo, TR (alumno) accede a todos los términos relacionados con alumno. KWIC (alumno) obtiene todos los términos que contengan la palabra "alumno" (índice permutado);
v) obtener todos los términos pertenecientes a una categoría dada, organizados con las estructuras semánticas que estén definidas en esa categoría; y
vi) obtener la macroestructura.

En cuanto al uso de las reglas de integridad:

i) si el modelo de datos dispone de reglas para preservar la integridad en las modificaciones, se deben aplicar para definir los cinco tipos de modificaciones previstas para los tesauros (ver capítulo 4, sección 4.3.3). En caso contrario se deben estudiar y definir en los algoritmos de modificación los procedimientos para mantener la integridad del tesauro; y

ii) si el tesauro es de indexación o explotación, los términos modificados o borrados deben mantenerse en un histórico con una referencia de tipo USE ó TG/TE a algún término del tesauro. Sólo pueden eliminarse del histórico cuando se pueda garantizar que los términos borrados no están siendo utilizados en alguna indexación.

Los modelos de datos informáticos, además, proporcionan una metodología de diseño basada en la independencia entre los datos, es decir, del contenido, estructura y presentación, de forma que las modificaciones hechas en un nivel, por ejemplo en la presentación, no afectan, o afectan de forma controlada, a los otros niveles (Ullman, 1988). Esto facilita el mantenimiento y gestión de los tesauros, que es una de las cuestiones más costosas en los tesauros no informatizados o construidos en formato digital pero sin aplicar un modelo de datos al diseño.

La *informatización* de los tesauros implica nuevas formas de concepción, terminológicos u ontológicos, y nuevos medios de construcción utilizando modelos y metodologías informáticas. Esto puede incrementar los costes, porque si esta actividad, era, inicialmente, abordada por documentalistas y lingüistas, ahora necesita la participación de especialistas en informática, convirtiéndose en una actividad interdisciplinar. Sin embargo, las ventajas y nuevas necesidades de los usuarios hacen imprescindible la informatización de los tesauros, porque, entre otras, se obtienen las siguientes ventajas (Aitchison y Clarke, 2004): (i) se automatiza la gestión del tesauro, mediante aplicaciones de carácter general o específico; (ii) se facilita la integración del tesauro en otros sistemas, por ejemplo, en sistemas RI o en herramientas *e-learning*; (iii) se facilita la integración y reutilización de diferentes tesauros; (iv) se facilita la interpretación y , en consecuencia, la compartición de los datos y la información entre sistemas informáticos y entre sistemas y personas; y (v) se mejora la efectividad y eficacia en el acceso a la información que contienen, creando más posibilidades de visualización, de exploración y de consulta. Esto implica una mayor probabilidad de que el usuario encuentre lo que busca y la ampliación de las estrategias de búsqueda automática del sistema RI o del sistema de gestión de recursos (Lancaster y Warner, 1993).

En este capítulo se revisan los modelos de datos utilizados para la construcción de tesauros informatizados, sean de la naturaleza que sean. En la figura 5.4 se presenta la clasificación de los modelos de datos aplicados al diseño de los tesauros que se va a utilizar para estructurar la revisión. Esta clasificación se hace tomando como base el

nivel de abstracción, que distingue entre los *modelos lógicos o conceptuales* y *modelos de implementación*[7]. Los primeros, modelos lógicos o conceptuales, se utilizan para analizar y representar el dominio, mientras que los segundos, los modelos de implementación, se usan para construir un sistema real. En el diseño de bases de datos se recomienda utilizar los modelos lógicos en una primera fase de diseño, y los modelos de implementación para traducir los esquemas lógicos a esquemas de datos directamente procesables por los sistemas de gestión de bases de datos. En definitiva, los modelos de datos lógicos describen la organización de los datos, la información o el contenido de forma cercana a las personas, mientras que los modelos de implementación ofrecen descripciones menos comprensibles para las personas y más cercanas a las máquinas. Existen correspondencias entre ambos tipos de modelos, aunque no siempre los modelos de implementación son capaces de reproducir los diseños lógicos.

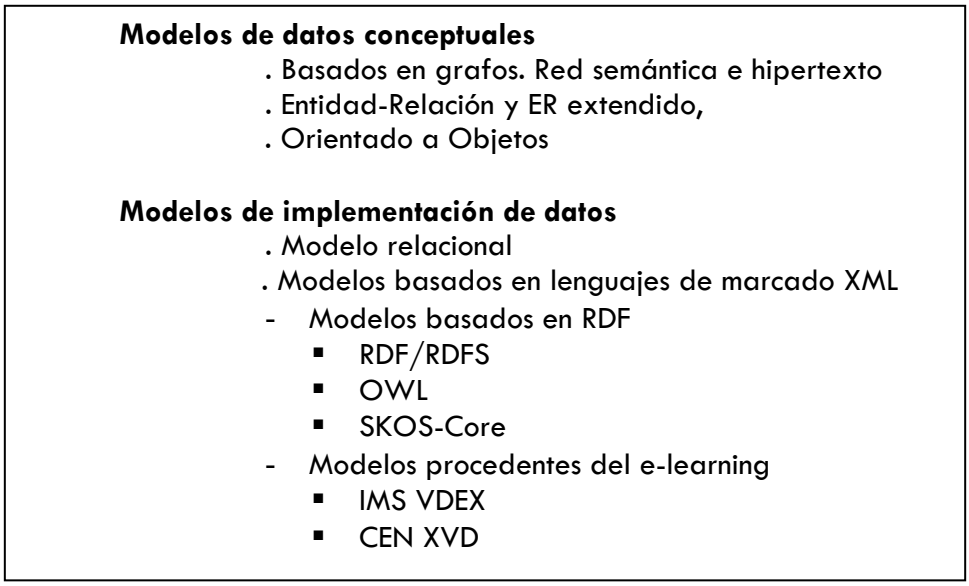


Figura 5.4. Clasificación de los modelos de datos para la informatización de los tesauros

5.2. Modelos de datos conceptuales

5.2.1. Modelos basados en grafos

Un grafo es un conjunto finito y no vacío de nodos y un conjunto de arcos, que se representan por pares de nodos, que representan las relaciones binarias entre nodos adyacentes. Matemáticamente se define como un par de conjuntos N, nodos y A, arcos:

$$G=(N,A);\ N=\{n_1, n_2,\ldots n_n\};\ A \subseteq N\times N$$

[7] Ver capítulo 4, sección 4.1.

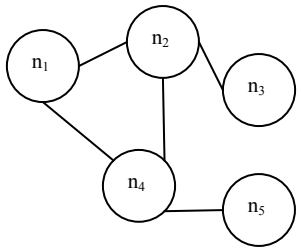


Figura 5.5. Diagrama de un grafo ejemplo

En la figura 5.5 se muestra el grafo G: N = $\{n_1, n_2, n_3, n_4, n_5\}$ y A = $\{(n_1, n_2), (n_1, n_4), (n_2, n_4), (n_2, n_3), (n_4, n_5)\}$.

Un grafo etiquetado es un grafo en el que los arcos tienen asociada una etiqueta que indica algún valor asociado al arco (figura 5.6):

$$G=(N,A,E);\ N=\{n_1, n_2,\ldots n_n\};\ A \subseteq N\times N\times E$$

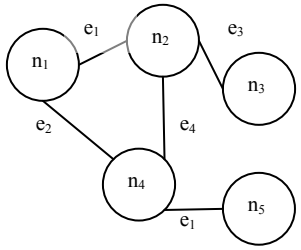


Figura 5.6. Diagrama de grafo etiquetado

Un grafo dirigido es un grafo donde los arcos tienen definido un sentido, lo que se representa con pares de nodos ordenados. Los grafos se representan, gráficamente, con diagramas y son una herramienta de diseño conceptual flexible y especialmente adecuada para modelar dominios con gran cantidad de interrelaciones, como el caso del dominio léxico (Fernández-Pampillón et. al, 2003).

En teoría de grafos e informática están definidos los procedimientos y algoritmos para recorrer los grafos, buscar la información que contienen y modificarla de forma consistente (Hortalá, et. al. 2001). Todas las operaciones básicas -enumeradas en la sección anterior- para acceder y manipular el contenido de los tesauros, están basadas en recorrer y buscar nodos en un grafo. En consecuencia, el contenido de un tesauro puede representarse con grafos (figura 5.3) considerando que los nodos representan términos y categorías, los signos lingüísticos del sistema, y los arcos representan las relaciones

semánticas y los otros campos que pueden formar parte de la microestructura[8] (figura 5.7). El tipo de relación semántica se representa con una etiqueta. La dirección de la relación se representa con un arco dirigido. El resultado es una representación del tesauro como un grafo dirigido y etiquetado:

Tesauro=(N,A,E); N={ti | ti es un término del tesauro};

E = {TG, TE, USE, USE FOR, TR,NA}

$A \subseteq N \times N \times E$

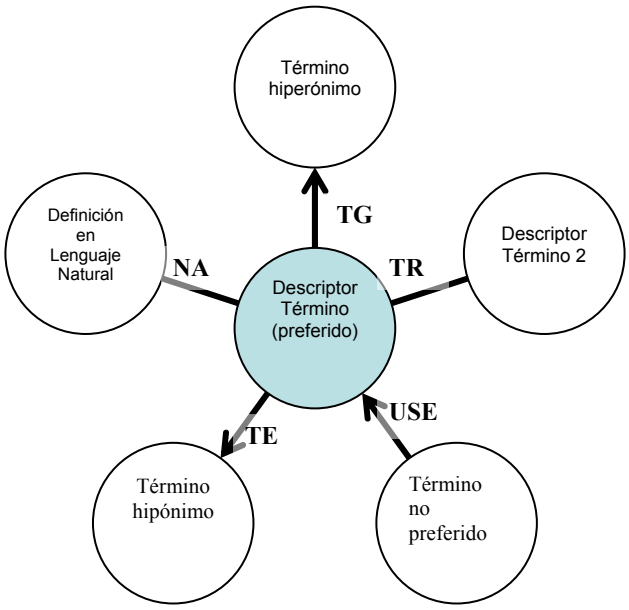


Figura 5.7. Diagrama de microestructura

Las relaciones del tesauro descomponen el grafo en tres tipos de subgrafos[9]:

1) árboles, respecto TE, que son grafos acíclicos, de generalización/especialización del dominio[10];
2) redes de familias semánticas, respecto TR; y,
3) clases de equivalencia, respecto de USE.

Sin embargo, a pesar de esta versatilidad, los grafos tienen ciertas limitaciones de representación que afectan a los tesauros. En primer lugar, no disponen de un mecanismo de definición de conjuntos de nodos, lo que significa que para representar las categorías se necesitan mecanismos adicionales como la definición artificial de la

[8] Por ejemplo, las notas de ámbito.

[9] G' = (N', A') se denomina subgrafo de un grafo G = (N, A), si G' es también un grafo y $N' \subseteq N$ y $A' \subseteq A$.

[10] En este sentido, resulta útil que el tesauro tenga marcados los términos "raíz" de la jerarquía (TT: Top Term).

relación categoría (Yan et al., 2006) o de la relación de pertenencia a una clase (IS-A) (Jones, 1993) (figura 5.8).

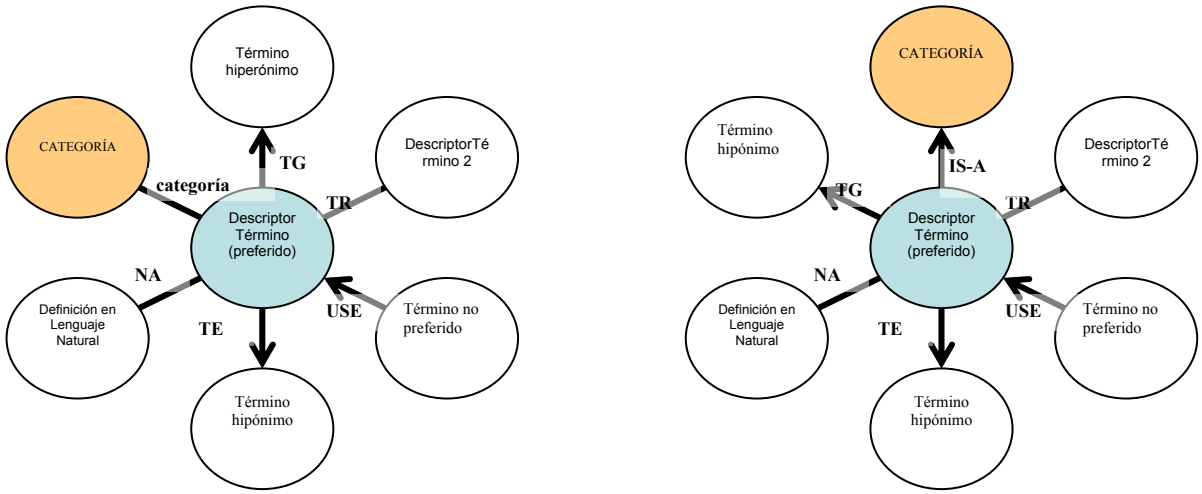


Figura 5.8. Posibles representaciones de la pertenencia de términos a una categoría en un tesauro (etiquetas "categoría" o "IS-A").

En segundo lugar, sólo pueden definirse relaciones binarias entre dos términos y en los tesauros existen relaciones asociativas (TR) entre más de dos términos en los que participan todos con todos, cuyo alcance no queda bien definido si se utilizan relaciones binarias[11]. La figura 5.9 muestra un ejemplo extraído del tesauro EUROVOC[12]: los términos incendio, 'lucha contra incendios', 'producto inflamable' y 'protección del bosque' son Términos Relacionados (TR) pero, en el grafo, sólo se definen relaciones entre ellos y el término 'lucha contra incendios'. La información, en consecuencia, no es completa. Existe, sin embargo, la posibilidad de extender el modelo de grafos al modelo de hipergrafos[13] que generaliza los arcos a hiperarcos para enlazar dos o más nodos (figura 5.10).

[11] De hecho, en la mayoría de los tesauros no se incluye esta información. Cuando dos o más términos están relacionados con un tercero mediante asociatividad (TR) la relación asociativa entre los primeros no se representa, sólo las relaciones entre el primero y el tercero y entre el segundo y el tercero.
[12] Tesauro plurilingüe que indexa los documentos de todos los ámbitos de actividad de la Comunidad Europea. Disponible en: http://europa.eu/eurovoc/
[13] Los hipergrafos son grafos con arcos e hiperarcos. Un hiperarco es una relación entre dos o más nodos que forma un conjunto de nodos relacionados todos con todos.

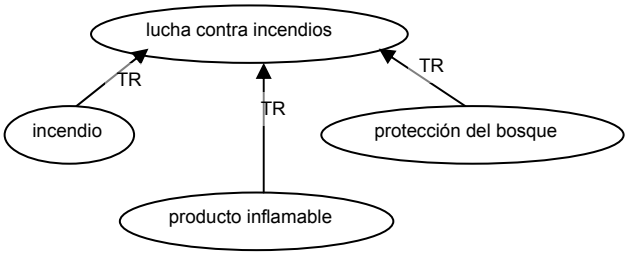


Figura 5.9. Representación, con grafos, de la relación asociativa (TR) entre cuatro términos del tesauro EUROVOC (en la microestructura del término 'lucha contra incendios')

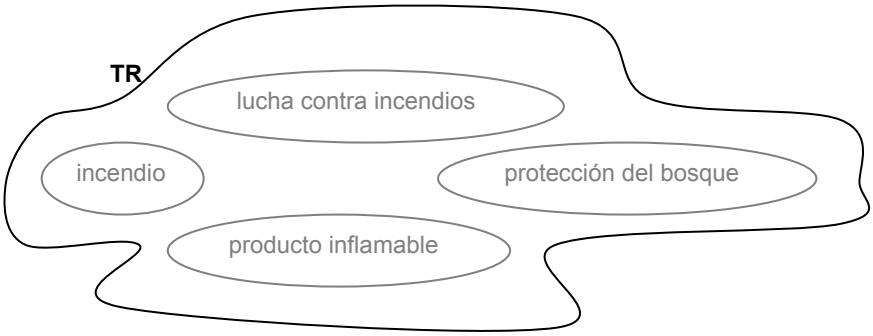


Figura 5.10. Representación de un hiperarco TR

En tercer lugar, no es posible definir relaciones entre conjuntos de nodos y/o nodos. En el tesauro podemos encontrar el caso de que las categorías estén relacionadas con términos[14]. Asimismo, la relación USE define la relación entre un conjunto de términos equivalentes, por ejemplo sinónimos, y un término preferido[15]. Al igual que en el caso anterior, en un grafo esta relación se representa mediante las relaciones binarias con el término preferido, perdiéndose la relación de equivalencia entre ellos. Por ejemplo, si escogemos de nuevo el tesauro EUROVOC, el término 'abandono escolar' lo encontramos referidos tres sinónimos:

abandono escolar

- UP abandono de la escolaridad
- UP abandono de los estudios
- UP interrupción de los estudios

[14] Ver el segundo caso práctico del capítulo 8.
[15] Esta estructura es similar a la estructura de 'synset' de la base de datos léxica WordNet (Miller, 1995), donde todos los términos sinónimos respecto de un significado se agrupan en un conjunto. Cada synset es un concepto.

Si buscamos 'abandono de la escolaridad', encontramos la relación de equivalencia con 'abandono escolar', pero no con el resto de los sinónimos 'abandono de estudios' e 'interrupción de los estudios'.

En el capítulo siguiente presentamos un modelo que extiende los grafos e hipergrafos y que, aplicada al dominio léxico, permite resolver estas limitaciones respecto a la representación de los tesauros.

Como muestra del uso de grafos en la construcción de tesauros se puede consultar el tesauro VisualThesaurus[16], un tesauro multilingüe con una interfaz que permite visualizar la parte del diagrama del tesauro relacionada con un término de consulta. Finalmente, es interesante destacar el hecho de que los grafos son el formalismo de representación subyacente de otros modelos de datos que, simplemente, incorporan semántica a los elementos estructurales de los grafos. De esta forma pueden definirse operaciones a nivel sintáctico y semántico y obtener un nivel de procesamiento más inteligente. Entre los modelos de datos basados en grafos aplicados a la construcción de tesauros destacan las redes semánticas y el hipertexto.

5.2.1.1. Redes Semánticas

Una red semántica es un tipo de grafo ampliamente utilizado para la representación del conocimiento en Inteligencia Artificial (Quillan, 1968; Sowa, 1991). En este modelo los nodos representan conceptos, instancias[17] o valores de atributos, y los arcos representan relaciones semánticas o atributos. La relación semántica más importante es la relación ES-UN (IS-A)[18], sobre la que funciona el mecanismo de inferencia por herencia[19].

Un tesauro puede considerarse una red semántica pero con matices: (1) los términos no son estrictamente hablando, conceptos, ni clases de objetos (Schauble, 1987); (2) la relación TG/TE no siempre es una relación de tipo/subtipo (ES-UN) sobre la que implementar inferencias por herencia[20] (Brachman, 1983); y (3) la riqueza de relaciones en una red semántica no es comparable con la simplicidad del tesauro (Gilchrist, 2003).

---

[16] Disponible en: http://www.visualthesaurus.com

[17] Una instancia de un concepto es un objeto o ejemplar concreto de ese concepto.

[18] Esta relación no siempre tiene el mismo significado que la relación jerárquica TG/TE (BT/NT) de los tesauros, a pesar de que ambas generen jerarquías de generalización/especialización (Brachman, 1983).

[19] En una red semántica con herencia simple la información o los datos asociados a un nodo se calculan con la unión de la información o datos propios de ese nodo y el de sus antecesores en la relación ES-UN o en la relación INSTANCIA.

[20] Las relaciones de hiperonimia-hiponimia entre términos son semánticas no de herencia, porque los atributos de los términos, notas de ámbito, notas históricas o código, no son heredables.

En cualquier caso, las redes semánticas se utilizan para diseñar los tesauros a nivel conceptual (Yan et. al., 2006), incluso a nivel conceptual y de implementación de datos (UMLS, 2008).

Entre los ejemplos de tesauros que aplican el modelo de redes semánticas destaca, por su riqueza y complejidad semántica y estructural, el Sistema de Lenguaje Médico Unificado (Unified Medical Language System), UMLS (UMLS, 2009). Este macrotesauro organiza las categorías y los términos en dos partes diferenciadas y conectadas. Las categorías se denominan Semantic Types y forman una red semántica con 135 categorías conectadas con 54 relaciones semánticas[21] (figura 5.11). Los términos forman en el llamado metatesauro[22.] y están conectados con las relaciones semánticas estándar de los tesauros: jerárquicas (BT/NT), asociativas (RT), y de equivalencia (USE/USE FOR); cada término pertenece como mínimo a una categoría y siempre a la categoría más específica posible.

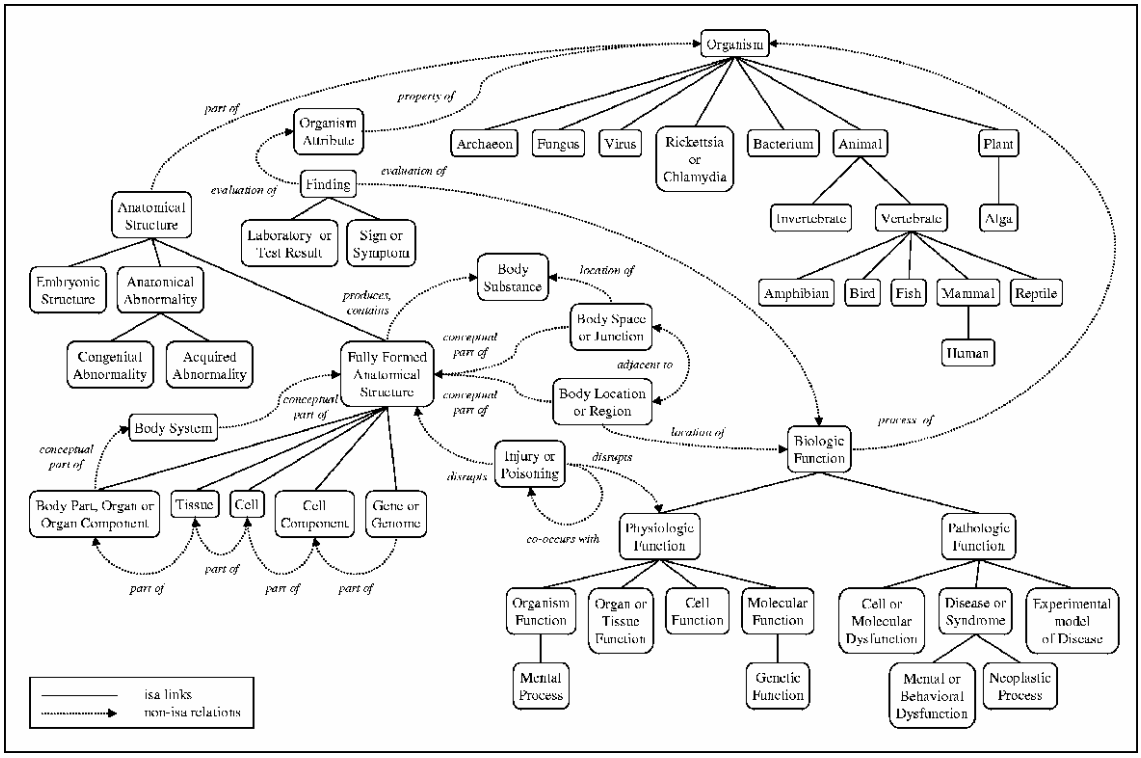


Figura 5.11. Parte de la Red Semántica UMLS (fuente (McCray, 2003))

La red semántica se ha implementado con una base de datos relacional[23] y con un archivo de registros de texto ASCII[24].

[21] Se clasifican en seis tipos: ES-UN (IS-A), "physically related to", "spatially related to", "temporally related to", "functionally related to", y "conceptually related to".
[22] El término metatesauro denota un tesauro más general capaz de comprender y trascender otros tesauros.
[23] El modelo relacional se revisará en la sección 5.3.1.

El esquema de datos relacional contiene en seis tablas:

1) dos tablas básicas, SRDEF con la definición de las categorías y relaciones semánticas y SRSTR con la definición de la estructura IS-A de la red;
2) dos tablas complemento, SRSTRE1 y SRSTRE2 con la definición del resto de las estructuras generadas por las relaciones no jerárquicas en un formato de enlace binario: categoría1, relación, categoría2;
3) dos tablas de administración, con información administrativa de los archivos de las tablas en SRFIL, y sobre los campos SRFLD.

El archivo de registros (US) contiene una secuencia de registros de dos tipos: de categoría y de relación. Estos registros se organizan en pares de campo-valor, donde el campo es una marca predefinida. Por ejemplo, UI: identificador único de la categoría; STY: nombre de la categoría, STN: número jerárquico de la categoría o DEF: definición de la categoría (figura 5.12).

```
UI: T020
STY: Acquired Abnormality
STN: A1.2.2.2
DEF: An abnormal structure, or one that is abnormal in size
or location, found in or deriving from a previously normal
structure. Acquired abnormalities are distinguished from
diseases even though they may result in pathological
functioning (e.g., "hernias incarcerate").
```

Figura 5.12. Extracto del registro de la categoría Acquired Abnormality del archivo de texto US de la red semántica del UMLS.

El metatesauro es un vocabulario complejo, con varias finalidades y multilingüe, construido sobre una estructura conceptual única, almacenada en un archivo denominado MRCONSO.RRF; se ha creado a partir de las versiones electrónicas de varios tesauros y vocabularios del ámbito de la biomedicina y salud. Sus características fundamentales son dos (i) conserva los términos y relaciones de los vocabularios de origen, pero traducidos a un formato único (Rich Release Format), y (ii) cada término se relaciona con el concepto o conceptos de la estructura conceptual UMLS, según la interpretación que los expertos del dominio hacen del significado del término en el

[24] La utilización de dos sistemas de almacenamiento diferentes con la misma información no aparece justificada en el bibliografía de referencia. Probablemente el archivo de registros se genere a partir de la base de datos relacional por razones de legibilidad y eficiencia. Es difícil leer la red semántica a partir de las estructuras de datos, tuplas, de la base de datos; reconstruir un objeto complejo, como la microestructura de una categoría, implica varias operaciones de combinación de estructuras de datos, por lo que puede ser más legible y efectivo tener preparada esa información en otro archivo.

contexto del vocabulario de origen. Si un término aparece en distintas jerarquías de diversos vocabularios, el metatesauro mantiene esas diferencias, relacionándolo con conceptos diferentes e indicando a qué vocabularios de referencia pertenece cada interpretación. La conexión entre los términos del metatesauro y la red semántica de categorías se realiza a través de los conceptos. Cada concepto está asociado a una o varias categorías de la red semántica. De esta forma los conceptos están categorizados y, por ende, los términos asociados a cada concepto.

Aunque en esta sección únicamente nos interesa la red semántica, el esquema de datos del sistema completo UMLS constituye una muestra de cómo integrar vocabularios en un entorno común sin perder la especificidad terminológica y semántica de cada uno. Asimismo, merece destacarse cómo se resuelve la representación de estructuras combinadas de conjuntos, categorías, y grafos, términos y relaciones semánticas: representando la relación de inclusión de los conjuntos con una red semántica IS-A y estableciendo dos correspondencias: la red semántica categorías con la red conceptual y la red conceptual con la red terminológica.

La representación de un tesauro como una red semántica tiene la ventaja de reflejar la naturaleza relacional del tesauro utilizando el conjunto normalizado de relaciones semánticas. En consecuencia es posible: 1) construir aplicaciones generales de gestión y procesamiento de tesauros; 2) compartir y reutilizar el contenido de los tesauros en sistemas diferentes y para construir otros vocabularios, y 3) utilizarlo para una "recuperación inteligente"[25] de los objetos de contenido (López y Moreira, 2007). Estos tesauros modelados y utilizados como redes semánticas se denominan *tesauros conceptuales*. Con respecto a los tesauros conceptuales, existen trabajos de investigación que proponen la ampliación "controlada" de las relaciones semánticas para aumentar su eficacia, tratando de no perder ni las ventajas de la estandarización ni su simplicidad (Hazewinkel, 1997; Tudhope, et. al., 2001; López y Moreira, 2007).

Desde el punto de vista de manipulación, las redes semánticas tienen la ventaja de ser un modelo extensible con capacidad de cambiar su estructura y contenido sin tener que redefinir el esquema de datos. Sin embargo, las redes semánticas presentan algunas desventajas en la construcción de tesauros, entre las que destacamos: 1) no disponen de lenguajes de consulta y modificación generales que faciliten, a los usuarios, la gestión; 2) no es capaz de representar algunas restricciones de integridad de los tesauros[26]; 3) el

[25] Se refiere a la recuperación de información basada en un modelo cognitivo que es el tesauro.
[26] Por ejemplo, la obligatoriedad de un término preferido en un conjunto de términos equivalentes.

coste temporal de las operaciones de manipulación puede ser alto si la red es muy compleja, y finalmente; 4) no es posible representar, por ser un modelo basado en grafos, los conjuntos, las relaciones entre conjunto, las relaciones entre elementos y conjuntos, y los hiperenlaces.

5.2.1.2. El hipertexto

Un hipertexto es un grafo en el que los nodos son unidades textuales[27] de información que representan un único concepto o idea y los arcos son enlaces bidireccionales entre las unidades de información, con la posibilidad de etiquetarse para explicitar su significado.

La característica fundamental del hipertexto es que los enlaces son procesables por las máquinas (Balasubramanian, 1995). Además, los nodos pueden estar espacialmente distribuidos o residir en una misma localización, y deben tener un identificador único (URI) para referenciar a los nodos de los enlaces. Los sistemas hipertextuales son las herramientas para la gestión automatizada del hipertexto: creación, publicación, consulta/visualización y mantenimiento (Balasubramanian, 1995). En la actualidad, Internet, con la aplicación Web y la tecnología asociada (navegadores, lenguajes XML), constituye el soporte que ha permitido aplicar el modelo de hipertexto para la organización de la información y de recursos.

Aunque se puede definir el hipertexto, simplemente, como un grafo con arcos procesables, desde el punto de vista de diseño conceptual, es importante distinguir tres formas de hipertextos (Wang y Randa, 1998):

1) el hipertexto sin estructura: no existen restricciones respecto del tipo de nodos y enlaces. La Web es el ejemplo más significativo;
2) el hipertexto semiestructurado: constituido conforme a normas o recomendaciones sobre el tipo de nodos y enlaces que no son de obligado cumplimiento. El resultado es un hipertexto que no siempre se ajusta al esquema lógico recomendado. Por ejemplo, los blogs y wikis parten de un esquema estructural mínimo del gestor hipertextual, pero los usuarios pueden crear enlaces arbitrarios. Wikipedia, un sistema hipertextual de artículos enciclopédicos, con varios vocabularios controlados para su organización, sistema de categorías de Wikipedia, Sistema de Clasificación

[27] El hipertexto se denomina *hipermedia* cuando los nodos contienen, además de texto, cualquier otro tipo de formato de información: imagen, sonido o vídeo.

Universal y de UNESCO, y un esquema estructural basado en el sistema de gestión hipertextual wiki; y

3) el hipertexto estructurado, el hipertexto que conforma un esquema estructural predefinido, por ejemplo, los libros electrónicos. El tesauro hipertextual pertenece a este último tipo.

Esta distinción indica que el hipertexto estructurado, e incluso semi-estructurado, es un grafo con restricciones que conforma un esquema lógico. Para construir ese esquema se han propuesto varias aproximaciones: redes semánticas (Conklin, 1987), hipergrafos con semántica de navegación (Tompa, 1989), o vocabularios con relaciones semánticas, como las taxonomías, clasificaciones, tesauros, ontologías (Ashman y Simpson, 1999). Todas ellas comparten un principio básico: reproducir el modelo de organización de la memoria humana con el fin de facilitar la exploración de la información mediante navegación (Bush, 1945). Si, además, el hipertexto debe servir para la explotación de contenidos, el diseño de los esquemas lógicos cobra especial relevancia y deben ser cuidadosamente estudiados para que el usuario pueda navegar con sencillez, rapidez y coherencia. Esto evita la desorientación, uno de los principales problemas de los hipertextos poco estructurados (Protopsaltis y Bouki, 2005).

En la actualidad, la utilización del hipertexto en los tesauros en línea es muy frecuente, especialmente en los tesauros de indexación y búsqueda de información y de recursos. El objetivo de un tesauro hipertextual es ayudar al usuario a buscar lo que necesita, explorando mediante navegación en su contenido. Normalmente, estos tesauros presentan el esquema de navegación en los dos niveles de macroestructura y microestructura (figura 5.13)[28]:

1) navegación a nivel de macroestructura: se accede a la estructura general del tesauro desde la que se refina la búsqueda de conceptos generales hasta llegar al término o términos específicos. El usuario visualiza todo el dominio de información que modela el tesauro y lo explora seleccionando los caminos que deben llevarle al fin buscado; y
2) navegación a nivel de microestructura: se accede directamente al término buscado. Es un procedimiento adecuado para la búsqueda focalizada, en la que el usuario selecciona el término que necesita y, desde su

[28] En RI se aplican, además de la navegación, otras técnicas más refinadas de búsqueda que guían al usuario. Estas técnicas están basadas en asignar pesos a los enlaces para indicar la mayor o menor probabilidad de escoger un camino u otro en función de criterios como la frecuencia de uso, la cercanía semántica de los términos enlazados, etc. (Jones, et. al, 1995).

microestructura, puede seguir navegando, seleccionando otros términos relacionados.

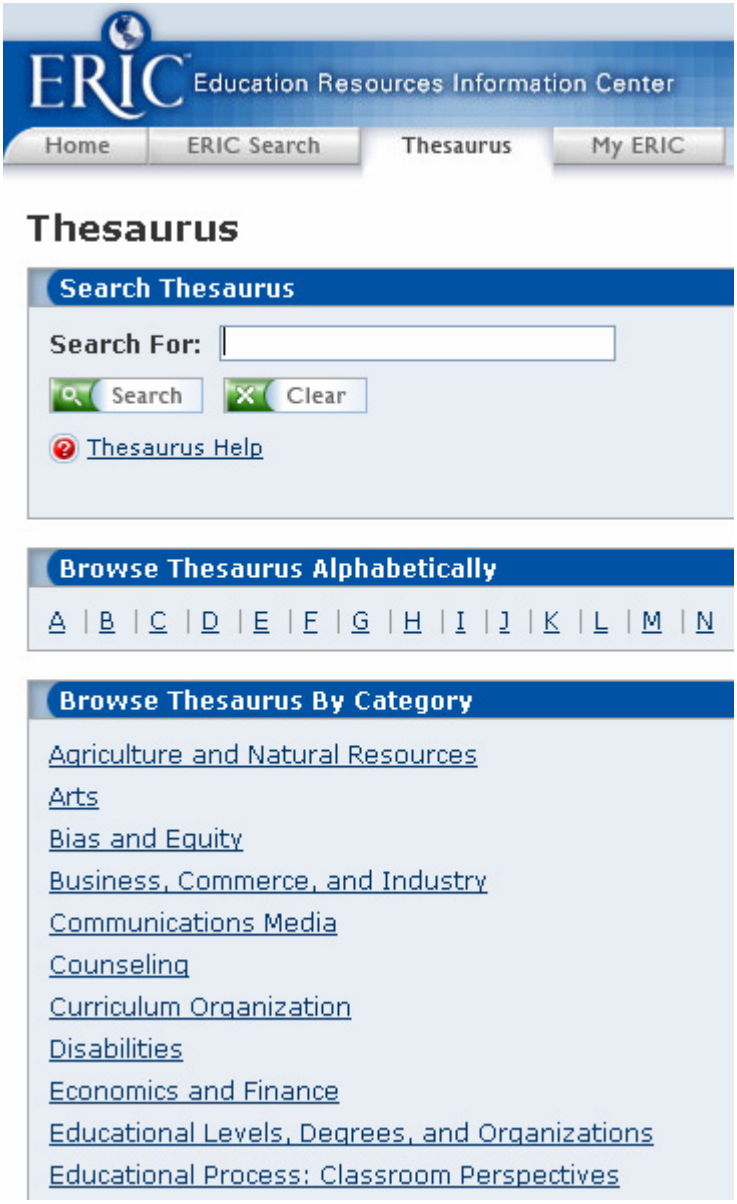


Figura 5.13. Esquemas de navegación hipertextual: focalizado, mediante la consulta y exploración alfabética, y sistemático, mediante el esquema de categorías principales del dominio

Estos esquemas de navegación reproducen los esquemas tradicionales de presentación de los tesauros: el alfabético y el sistemático, y no en todos los casos se corresponden con los esquemas de datos de un tesauro hipertextual. Nos encontramos con dos situaciones posibles: 1) una presentación hipertextual añadida al tesauro para visualizar y consultar el contenido. Un ejemplo en línea de este tipo de presentación hipertextual

son los tesauros del CINDOC[29]; y 2) una presentación hipertextual de un tesauro hipertextual, en la que el usuario ve y utiliza el propio esquema conceptual del tesauro para navegar (Duncan, 1990; Pollard, 1993). Los tesauros construidos automáticamente suelen utilizar estos esquemas hipertextuales (Huggett y Lanir, 2007).

En resumen, el hipertexto es una red de nodos complejos que puede estar estructurada. Esta estructura se diseña a partir de otros modelos de datos, habitualmente redes semánticas. El tesauro hipertextual es un tipo de hipertexto estructurado que aporta, frente a otros modelos, la posibilidad de que las personas puedan explorar el dominio conceptual de la información y los objetos indexados por el tesauro para alcanzar una idea clara del contenido de esta información y objetos y, en consecuencia, buscar y seleccionar con facilidad lo que necesitan. Desde el punto de vista de recuperación de información, el tesauro hipertextual puede entenderse como una herramienta puente entre los intérpretes de consultas en lenguaje natural y los mapas conceptuales u ontologías de un dominio de conocimiento particular (Jones et al., 1995).

5.2.2. Modelos Entidad-Relación y Entidad-Relación Extendido

El modelo Entidad-Relación (ER) es un modelo de datos visual orientado fundamentalmente al diseño de bases de datos *relacionales*[30] (Chen, 1976). Este modelo ha sido, posteriormente, extendido y modificado[31], dando lugar al modelo Extendido Entidad Relación (EER) que es, en realidad, una familia de modelos con la forma gráfica común de representación del modelo ER (Teorey et. al., 1986; Elsmari y Navathe, 1997). Estos modelos son adecuados para el diseño de bases de datos que utilizan procesos de diseño descendentes, de un esquema general a esquemas detallados, para reducir la complejidad y minimizar el riesgo de errores. Los esquemas de datos que se obtienen se denominan diagramas ER (o EER) y se pueden transformar, directamente, en esquemas de datos relacionales normalizados[32].

Los constructores básicos del modelo son las entidades, relaciones y atributos (figura 5.14).

---

[29] Disponible en: http://thes.cindoc.csic.es/

[30] Aquí la palabra relacional se refiere al concepto matemático de relación. Posteriormente presentaremos este tipo de modelo de datos de implementación de bases de datos.

[31] Para introducir restricciones semánticas, restricciones de integridad, relaciones de generalización y herencia, atributos compuestos, relaciones temporales, etc.

[32] Estas transformaciones pueden ser automáticas cuando se utilizan herramientas CASE de apoyo al diseño de bases de datos. Los esquemas relacionales se dice que están normalizados si cumplen una serie de condiciones que aseguran la mínima redundancia, la integridad y el cumplimiento de restricciones previamente definidas.

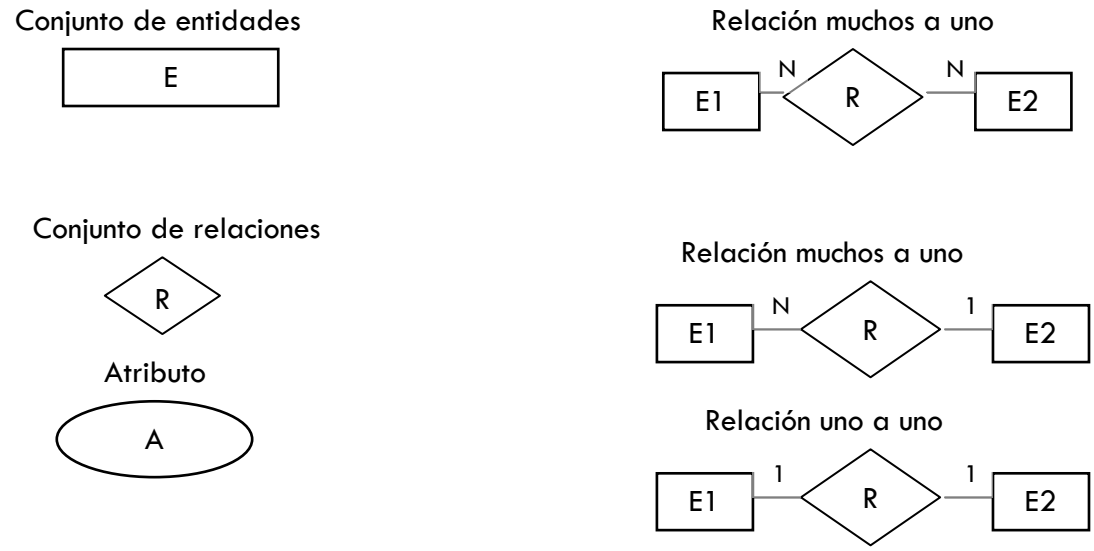


Figura 5.14. Representación gráfica de los constructores básicos (izquierda) y las restricciones de cardinalidad (derecha)

Las *entidades* son todo aquello que tiene existencia física o abstracta. Se clasifican en *tipos de entidades* mediante un predicado que define la pertenencia de una entidad al tipo. Por ejemplo, cada término de un tesauro es una entidad que pertenece al tipo de entidades TÉRMINO (figura 5.15). Los tipos de entidades se representan gráficamente con un rectángulo con una etiqueta en su interior con el nombre del conjunto.

Las *relaciones o interrelaciones*[33] son asociaciones entre entidades. Un tipo de relaciones es una relación matemática entre n entidades de n tipos de entidades: {(e1, e2, . . , en | e1 ∈ El, e2 ∈ E2, . . . , en ∈ En}.

Cada tupla (e1, e2, . . . , en) es una interrelación. Puede ocurrir que las entidades que se relacionan sean de un mismo tipo, como ocurre con las relaciones del tesauro entre entidades del tipo TÉRMINO. En este caso, es posible definir los papeles con los que participa cada entidad en una relación. Por ejemplo, la relación USE entre términos debe especificar qué tipo de términos tiene el papel de *preferido* y cuáles son *no preferidos*. Por ejemplo, la relación (interrelación) USE en el tesauro Spines[34] incluye una marca en cada término:

$USE_{tesauro\ Spines}$ = {(preferido/ABONO, no preferido/FERTILIZANTES ORGÁNICOS, (preferido/ABONO, no preferido/ABONO DE CORRAL, …}

[33] Utilizamos el término interrelación para el término relación cuando significa asociación y queremos desambiguarlo del sentido matemático de relación, que es un conjunto de tuplas.

[34] Tesauro del CINDOC de ámbito científico y técnico. Disponible en: http://thes.cindoc.csic.es/index_SPIN_esp.html

La definición de un conjunto o tipo de relación incluye la definición de sus restricciones: la aridad y la cardinalidad. La *aridad* de un tipo de relación es el número de tipos de entidades que participan en la asociación, y es el tamaño de las tuplas del conjunto. La *cardinalidad* especifica el número máximo de entidades de un conjunto que pueden relacionarse con entidades de otro conjunto. Se distinguen tres (figura 5.14):

1) muchos a muchos: una entidad de un conjunto de entidades E1, puede relacionarse con cero, una o varias entidades de E2 y viceversa. Por ejemplo la relación de tipo TR de la figura 5.15;
2) muchos a uno o uno a muchos: una entidad de un conjunto de entidades E1 puede relacionarse como máximo con una entidad del conjunto E2, pero una entidad del conjunto E2 puede relacionarse con varias entidades de E1. Por ejemplo, TG y USE de la figura 5.15; y
3) uno a uno, una entidad del conjunto E1 puede relacionarse como máximo con una entidad del conjunto E2, y viceversa, una entidad de E2 puede asociarse como máximo con una de E1.

Los *atributos* son las características o propiedades de un conjunto de entidades. Cada entidad tiene un valor o una tupla de valores para ese atributo que lo distinguen de otras entidades del conjunto. Un atributo o grupo de atributos se define como *clave*, y se subraya en el diagrama, si su valor es único para cada entidad, por lo que sirve para identificarlas. Formalmente, es una función entre un conjunto de entidades o relaciones y un conjunto de valores o el producto cartesiano de un conjunto de valores:

$$A: Ei \text{ or } Ri \rightarrow Vi \text{ or } Vi, X\ Vi, X \ldots X\ Vi,.$$

Por ejemplo, el diagrama de la figura 5.15 muestra que el 'descriptor', 'código' y 'nota de ámbito' son atributos de los términos. El descriptor es un atributo clave[35]. Así, el término 'abandono escolar' se define con los valores de este conjunto de atributos:

DESCRIPTOR= ABANDONO ESCOLAR;
NA=ÚSESE EN RELACIÓN CON ESTUDIANTES QUE, POR INEPTITUD, RAZONES ECONÓMICAS U OTRAS, NO LLEGAN A COMPLETAR LOS ESTUDIOS DE UN CICLO O PERIODO DETERMINADO;
CÓDIGO = null[36]

[35] Los atributos clave identifican de forma única a cada término.
[36] La constante null se utiliza en bases de datos para indicar que no tiene valor, bien porque no existe o porque no se conoce.

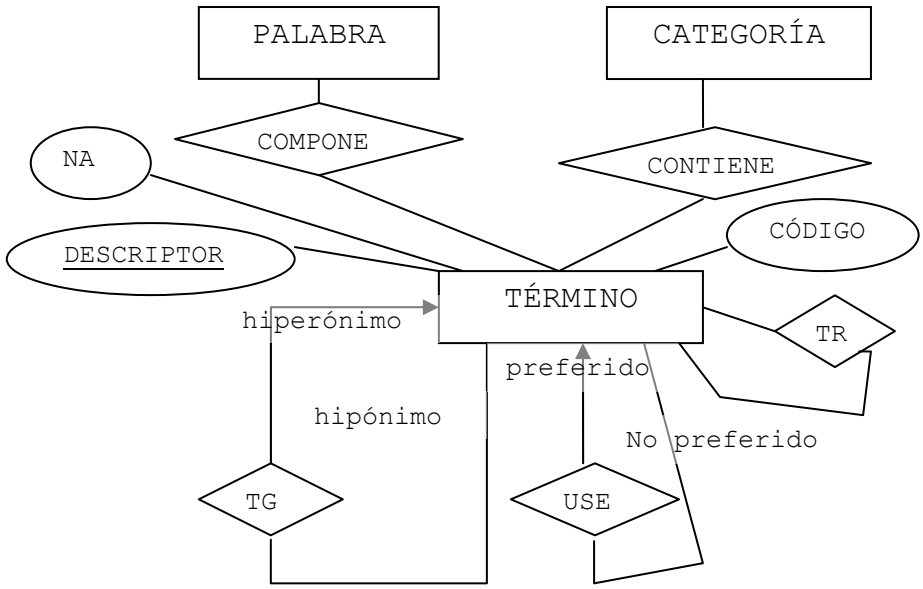


Figura 5.15. Diagrama ER para la representación del contenido de los tesauros monolingües conforme al modelo estándar (UNE 50106, 1990)

El diagrama ER de la figura 5.15 representa las siguientes informaciones del modelo estándar para tesauros monolingües (UNE 50106, 1990):

1) un tesauro monolingüe es un conjunto de cero, una o más categorías;
2) las categorías contienen cero o más términos. Un término está contenido en cero o más categorías;
3) los términos se caracterizan por los valores de sus atributos: descriptor, código y NA (nota de ámbito). El valor del atributo descriptor es único y obligatorio para cada término;
4) un término se compone de una o más palabras. Una palabra puede formar parte de uno o más términos;
5) un término, con el papel de hipónimo, puede tener como máximo un único término hiperónimo[37]. Un hiperónimo puede tener cero, uno o más hipónimos; y
6) la relación de equivalencia tiene los papeles de: término no preferido y término preferido. Un término preferido se relaciona con cero, uno o varios términos no preferidos, cada término no preferido se relaciona con cero o un término preferido.

[37] En este diagrama, esta restricción prohíbe explícitamente jerarquías múltiples. Si la cardinalidad fuera muchos a muchos las jerarquías de hiperonimia-hiponimia podrían ser múltiples.

7) Cada término puede estar relacionado por asociatividad con cero o más término.

Sin embargo no es posible representar:

1) las clasificaciones: los términos se clasifican en preferidos y no preferidos, porque sólo los preferidos se relacionan por TG (hiperonimia) y TR (asociatividad), los no preferidos se relacionan por USE con los preferidos (figura 5.16). Además, sólo los términos preferidos tienen atributos de código y notas de ámbito;
2) la participación total o parcial de las entidades en las relaciones. Por ejemplo, respecto de la relación TG, la participación de hiperónimos e hipónimos es total: si un término es hipónimo tiene siempre un hiperónimo y viceversa. Respecto de la relación USE, un término no preferido tiene siempre un término preferido (total) pero no es cierta la inversa (parcial);
3) las restricciones de cardinalidad de las entidades máxima y mínima, están muy relacionadas con la participación. Los términos preferidos se relacionan por equivalencia, como mínimo con cero términos no preferidos (participación parcial) y como máximo con un número cualquiera (0,N); los términos no preferidos se relacionan (USE) como mínimo y máximo con un término preferido (1,1) lo que se corresponde con una participación total;
4) las relaciones N-arias de equivalencia y asociatividad (hiperarcos);
5) la posibilidad de distinguir subtipos en las relaciones: la relación general/específico (TG/TE) podría ser hiperonimia-hiponimia, meronimia-holonimia, …; y
6) conjuntos de conjuntos. El conjunto CATEGORÍA debería contener al conjunto TÉRMINO. En los diagramas se representa una relación conjuntista de inclusión con una relación binaria entre los conjuntos TERMINO Y CATEGORÍA.

Las tres primeras limitaciones pueden resolverse utilizando el modelo EER. Las extensiones necesarias son las de tipo/subtipo, parcialidad/totalidad y cardinalidad del conjunto de entidades (figura 5.16). Sin embargo las tres últimas limitaciones no se pueden representar con estos modelos ER y EER.

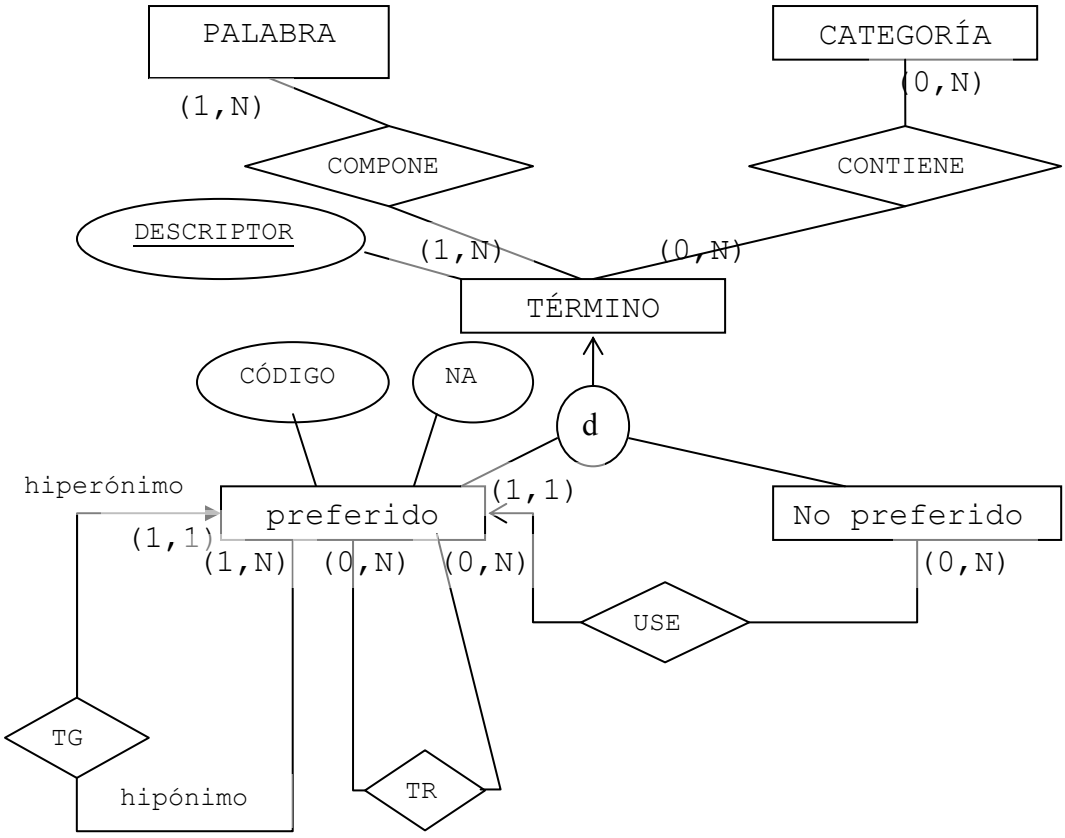


Figura 5.16. Diagrama EER para la representación del modelo estándar de tesauros monolingües (UNE 50106, 1990)

Los modelos ER y EER han sido utilizados para el diseño conceptual de tesauros almacenados en bases de datos relacionales (Martínez, et. al. 1992; Jones, 1993; Rodríguez, 1997). El tesauro *INSPEC* sobre Física, Tecnología eléctrica y electrónica, Computación, Ingeniería de control y Tecnología de información, está almacenado en una base de datos relacional modelada con ER (Jones, 1993) y es una de las herramientas para el acceso a la base de datos bibliográfica INSPEC[38]. El diagrama ER del tesauro INSPEC es, básicamente, el mostrado en la figura 5.15, con dos diferencias: la relación TG es muchos a muchos, porque el tesauro INSPEC es polijerárquico y la relación de equivalencia (USE) también. En este último caso, no se ha encontrado la razón de diseñar una relación uno a muchos como muchos a muchos.

Este modelo de datos no dispone de mecanismos para la definición de las consultas y modificaciones. Sin embargo, el esquema de datos debe representar todos los datos y relaciones que se necesiten para resolver las consultas previstas. Por ejemplo, si se

[38] Esta base de datos es accesible a través de ISI Web of Knowledge en: http://science.thomsonreuters.com/es/productos/inspec/

incluye una entidad para representar las palabras que componen los términos se garantiza que el tesauro se puede presentar como índice permutado.

El modelo ER y EER es adecuado cuando el tesauro se va a implementar con bases de datos relacionales porque 1) simplifican el diseño de los esquemas de datos de un dominio reduciendo la complejidad del dominio, al menos, en un factor de 10; 2) facilitan la definición de los tipos de relaciones y restricciones de integridad mediante la abstracción de las relaciones (Teorey et. al, 1986); además, 3) los diagramas se transforman directamente en esquemas normalizados de bases de datos relacionales. Sin embargo, como hemos visto, el modelo no tiene la capacidad expresiva suficiente como para representar la estructura de conjuntos e interrelaciones de los tesauros. Esto limita las operaciones y consultas que se pueden realizar sobre él[39] y, en el caso de tesauros de explotación, limita la búsqueda mediante la navegación por la estructura de términos[40].

5.2.3. Modelo Orientado a Objetos

El modelo de datos Orientados a Objetos (OO) forma parte de las metodologías para construir sistemas basados en la Programación Orientada a Objetos (Bertino et. al, 2001). La metodología estándar para el diseño de estos sistemas se denomina Lenguaje de Modelado Unificado, UML[41] (ISO/IEC 19501, 2005). Esta metodología, que ha sido desarrollada por el Grupo de Gestión de Objetos (OMG)[42], utiliza cinco tipos de diagramas para modelar los aspectos estáticos y dinámicos de sistemas software: diagramas de casos de uso, diagramas de estructura, diagramas de interacción, diagramas de estado, y diagramas de implementación (Harmon y Watson, 1998). Los diagramas que representan la estructura del sistema a nivel lógico son los diagramas de estructura y se construyen aplicando los constructores del modelo orientado a objetos (figura 5.17).

---

[39] Por ejemplo, si se consultan los términos relacionados TR con un término dado es posible que no obtengamos la respuesta si el término es no preferido, puesto que esta relación se especifica entre términos preferidos. Para localizarlos tendríamos que buscar (operación join) el término preferido (USE) y, con él, solicitar los términos relacionados, que además sólo serían preferidos; por lo tanto, no sería una respuesta completa.

[40] En este caso se añaden a los tesauros interfaces hipertextuales que reconstruyen la macroestructura del tesauro a partir de los datos del tesauro relacional (Jones et al., 1995).

[41] The Unified Modelling Language.

[42] OMG (http://www.omg.org) es un consorcio dedicado al cuidado y el establecimiento de diversos estándares de tecnologías orientadas a objetos, tales como UML, XMI, CORBA. Es una organización no lucrativa que promueve el uso de tecnología orientada a objetos mediante guías y especificaciones para tecnologías orientadas a objetos. El grupo está formado por compañías y organizaciones de software como HP, IBM, Sun Microsystems, Apple Computer (Wikipedia).

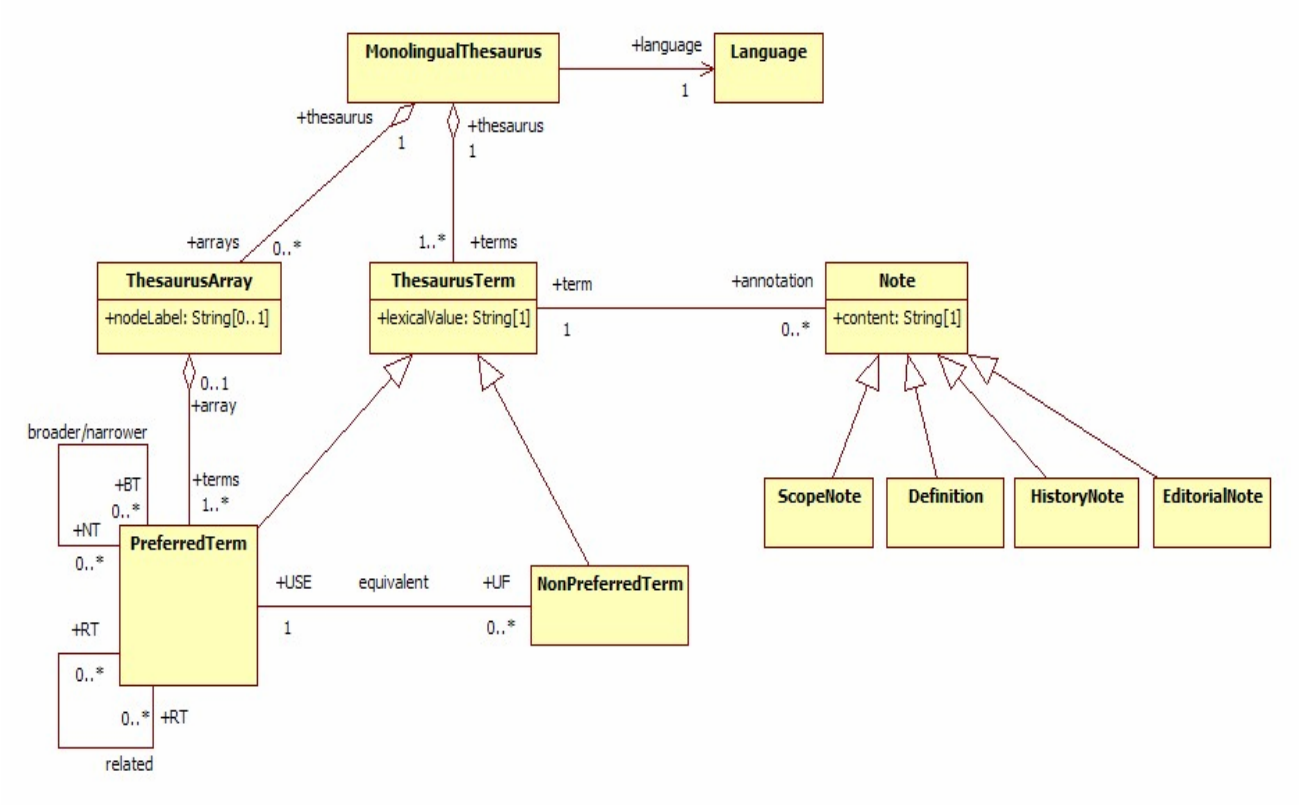


Figura 5.17. Diagrama de clases UML del estándar de tesauros BS 8723[43]

Los constructores básicos del modelo orientado a objetos son (Bertino y Martino, 1995):

1) *objetos:* cada una de las entidades reales o abstractas del mundo. Cada objeto se identifica de forma única con un identificador de objeto (OID);
2) *atributos (variables de instancia o slots)*: cada una de las propiedades que puede tener un objeto;
3) *objetos complejos*: objetos que contienen atributos con otros objetos como valores (objetos anidadados);
4) *métodos*: las operaciones entre los objetos. Los métodos tienen dos partes, el encabezamiento, que es la definición del método (nombre y tipos de objetos operadores) y el cuerpo, que es el código;
5) *encapsulamiento*: mecanismo para usar de forma transparente, sin ver el código, y de modo modular los métodos que contiene un objeto. Consiste en definir una interfaz en cada objeto con todas las operaciones que pueden aplicarse al objeto. Los valores de los atributos de los objetos se manipulan con estas operaciones definidas en la interfaz. De forma oculta o privada se define el código de cada método;

[43] Fuente (Miles, 2006).

6) *clase*s: tipos de objetos. Una clase agrupa y define todos los objetos que comparten el mismo conjunto de atributos y métodos; y
7) *herencia*: mecanismo de inferencia de conocimiento que se basa en la relación de especialización-generalización (IS-A) entre clases. La clase más específica (subclase) hereda todos los atributos y métodos de la clase más general (llamada superclase).

Un buen ejemplo de aplicación de estos conceptos al diseño de tesauros es el diagrama de la figura 5.18 (Wilson y Matthews, 2002). Este diagrama muestra un diseño de tesauro basado en conceptos, donde los términos se consideran propiedades de los conceptos, las realizaciones de los conceptos. El diagrama contiene las clases:

- ThesaurusObject: la clase raíz del tesauro. Cada tesauro es un objeto de esta clase;
- Concept: la clase a la que pertenecen cada uno de los conceptos del tesauro. Es un objeto complejo porque contiene como valores de alguna de sus propiedades otros objetos, por ejemplo la propiedad *indicated by* contiene objetos de la clase *Term* y *hasScopeNote*, objetos de la clase *ScopeNote;*
- TopConcept: subclase de la clase Concept que sirve para definir los conceptos que sean raíz de alguna jerarquía;
- Term: término para un concepto; y
- ScopeNote: notas de ámbito.

Las propiedades o atributos de los conceptos (clase Concept) son:

- ClassificationCode: identificador, no necesariamente único, de cada concepto;
- hasScopeNote: relación entre conceptos y las notas de ámbito, que son objetos de las clases Concept y ScopeNote respectivamente;
- isIndicatedBy: la relación entre conceptos y términos;
- PreferredTerm: es una subpropiedad de la propiedad isIndicatedBy que indica si el término es preferido;
- UsedFor: es la subpropiedad que indica si un término es no preferido; y
- ConceptRelation: es la generalización de todas las relaciones semánticas, puede tener como valores:
    - Broader,
    - Narrower,
    - Top concept
    - Related concept

Las propiedades de los Términos (clase Term) son:

- inLanguageOf: el lenguaje del término;
- value: forma ortográfica del término;

Las propiedades de la clase Scope Note son:

- inLanguageOf: lenguaje de la nota de ámbito; y
- ScopeNoteType: una de las subclases general, hierarchy, translation, editor o history

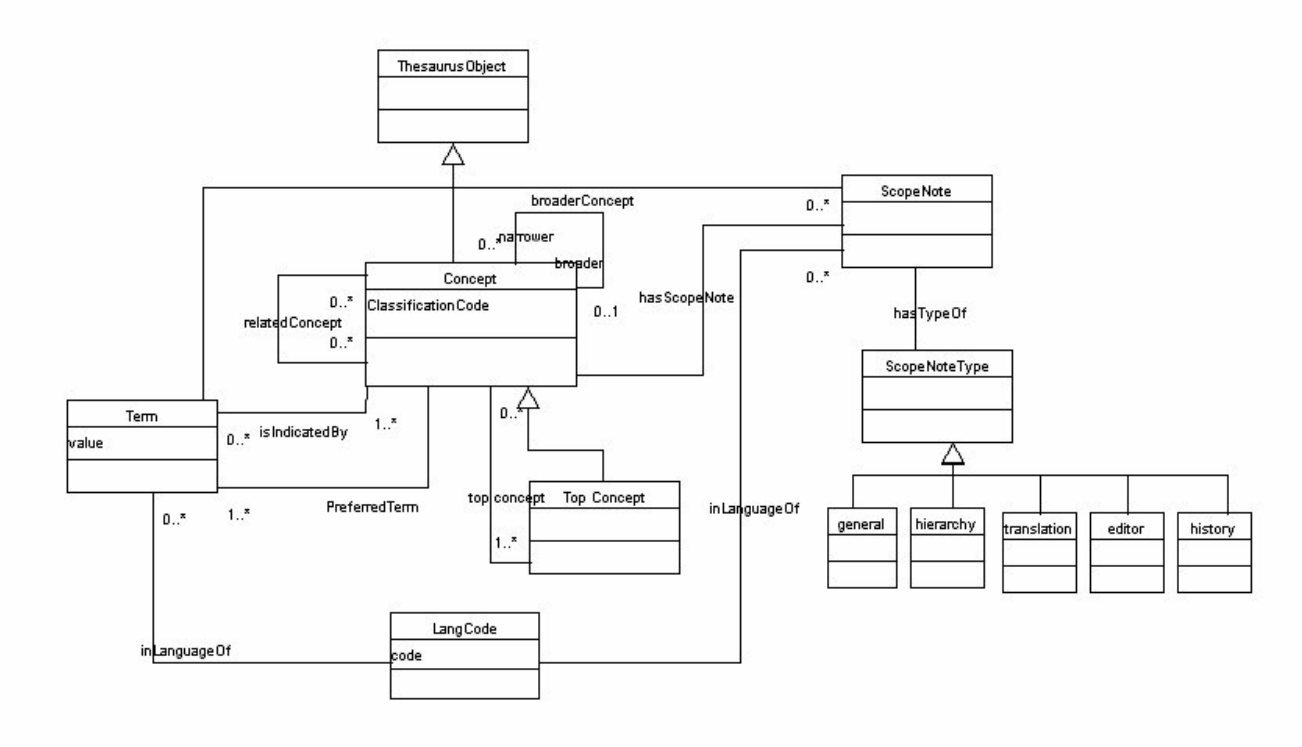


Figura 5.18. Diagrama de clases del modelo estándar de tesauro multilingue[44]

En Kerschberg and Weishar (2000) se presenta otra aplicación del modelo OO y la metodología UML para la construcción de tesauros, con el objetivo de que se integren con más facilidad en otros sistemas. En Liu, et al. (1999) se propone una metodología para la construcción de vocabularios controlados basada en el modelo OO, con el objetivo de mejorar, con estos vocabularios, la interoperabilidad entre aplicaciones software.

La aplicación del modelo Orientado a Objetos para la construcción de los tesauros tiene la ventaja de que el tesauro que resulta es fácilmente implementable e integrable en otros sistemas informáticos. Sin embargo, conviene tener en cuenta que es "un modelo de datos en un lenguaje de programación", por lo que el esquema de datos no es

[44] Fuente Wilson y Matthews, 2002.

independiente del lenguaje o del software que manipule el tesauro. En consecuencia, existen menos posibilidades de compartir y reutilizar el contenido del tesauro. También tiene el inconveniente de que no es un modelo cercano a las características del tesauro: las relaciones características de herencia e instancia no tienen sentido en el tesauro, lo que obliga a definir, de forma artificial, otra semántica a dichas relaciones (Génova, et. al, 2000).

5.3 Modelos de implementación de datos

Los modelos de implementación de datos son modelos cuyos esquemas de datos son directamente procesables por sistemas de gestión de datos, pero se mantienen en un nivel de abstracción superior al nivel físico que es el nivel del sistema de archivos. Estos modelos de datos son interpretados, almacenados y gestionados automáticamente por sistemas informáticos denominados Sistemas Gestores de Base de Datos (SGBD). El modelo más importante por su extensión de uso es el modelo relacional. Actualmente también está creciendo el uso de modelos basados en lenguajes de marcas XML, muy adecuados para el intercambio de información en el entorno de Internet. En este apartado revisamos ambas aproximaciones que, de hecho, pueden considerarse complementarias. Habitualmente las bases de datos se construyen con el modelo relacional pero intercambian datos con otras aplicaciones en formato XML.

5.3.1. Modelo relacional

El modelo relacional fue introducido por E. F. Codd en 1970 (Codd, 1970). Durante los diez años siguientes se desarrollaron los Sistemas de Gestión de Bases de Datos que permiten la manipulación automática de los datos. En los años 80 y 90 se generalizó su uso, especialmente en el mundo empresarial. Casi diez años después, Codd termina de completar el modelo añadiendo más operaciones, restricciones y extendiendo la semántica (Codd, 1979). Este trabajo de Codd es el que tomamos de base para hacer una breve introducción al modelo.

El modelo relacional consta de los siguientes tres componentes:

(1) un conjunto de relaciones;

(2) un lenguaje de datos con al menos los operadores del Álgebra Relacional; y

(3) un conjunto de reglas generales que preservan la integridad de los datos en las operaciones de inserción, actualización, borrado: integridad de entidades e integridad referencial.

Las *relaciones* son estructuras de datos formadas por una lista de valores pertenecientes a un conjunto de dominios. Un *dominio* es un conjunto de valores, por ejemplo, el dominio de los apellidos, es el conjunto de todos los apellidos que existen. Supongamos que se eligen *n* dominios D1, D2, …, Dn, no necesariamente distintos. El producto cartesiano de los *n* dominios D1 ×D2 × …× Dn es el conjunto de las n-tuplas (t1, t2, …,tn), tales que cada ti pertenece al dominio Di para todo i. Una relación de *grado n* es un subconjunto de tuplas del producto cartesiano de los dominios. Por ejemplo, la relación *alumnos* se define como un conjunto de tuplas del producto cartesiano de los dominios DNI×Nombre×Apellidos×Numero_Matrícula. Los nombres de los dominios son los atributos de una relación. Las relaciones pueden representarse en forma de tablas, con las cuatro siguientes propiedades (tabla 5.1):

1. no puede haber duplicación de filas (tuplas);
2. El orden de las filas es irrelevante;
3. El orden de las columnas es irrelevante; y
4. Los valores de cada celda deben ser atómicos.

DNI	Nombre	Apellido	Teléfono
2222	Pepe	Jiménez	91222222
3333	María	García	35889900
4444	Concha	Pérez	90000898

Tabla 5.1. Relación alumno formada por los atributos DNI, Nombre, Apellido y Num_Matrícula

Una *base de datos relacional* es una colección de relaciones, cuyos valores varían en el tiempo, son accesibles, y actualizables. Se puede visualizar como una colección de tablas. Entre las tablas no hay enlaces, pero están conectadas entre sí por los valores de ciertos atributos de conexión. Estas conexiones son las que van a permitir navegar entre los datos de las diferentes tablas de la base de datos. Por ejemplo, supongamos que se define, además de la tabla alumno (tabla 5.1), la tabla matriculas (tabla 5.2):

DNI	Asignatura
2222	Análisis del Discurso
2222	Sintaxis
3333	Sintaxis
4444	Lingüística Informática

Tabla 5.2. Relación Matrículas

Las tablas Alumnos y Matrículas están conectadas por el atributo DNI, de manera que las tuplas de una de las tablas están relacionadas con las tuplas de la otra cuando tienen el mismo valor en DNI. Por ejemplo, el alumno 2222, Pepe Jiménez de la tabla Alumnos, está matriculado en dos asignaturas: Análisis del Discurso y Sintaxis. Esta información se obtiene conectando la tupla 2222 de Alumnos, con todas las tuplas 2222 en Matrículas.

Las relaciones pueden tener definidas *claves candidatas*. Una de las claves candidatas se selecciona como principal se denomina *clave principal o primaria*. Una clave candidata es una colección de atributos de la relación que cumple dos condiciones:

1) unicidad: no pueden existir dos tuplas con la misma clave; y
2) mínimo número de atributos: si se elimina algún atributo de la clave se puede perder la unicidad.

Las claves permiten identificar de forma única cada tupla en una relación y son el mecanismo principal para mantener la consistencia aplicando las reglas de integridad.

Las reglas de integridad son restricciones que se imponen a las bases de datos relacionales para mantener la consistencia de los datos durante las operaciones de actualización de datos: inserciones, modificaciones y borrados. Las dos reglas básicas son:

1) regla de integridad de las entidades. Las claves primarias no pueden tener valores vacíos o nulos (null); y
2) regla de integridad referencial. Si en alguna relación de la base de datos R1, existe un valor v de un atributo A y A forma parte de una clave primaria de otra relación R2, entonces es necesario que, en la relación R2 denominada tabla madre, exista el valor v para A. Por ejemplo, el DNI en la relación Matrículas se refiere al DNI en la relación alumnos, que es la relación o tabla madre. La regla de integridad referencial obliga a que todos los valores del DNI de Matrículas estén definidos previamente en Alumnos. Por lo tanto, no podrían añadirse nuevas matrículas con DNI que previamente no estén definidos en la tabla Alumnos.

Finalmente, el último componente del modelo relacional es el *lenguaje de consulta*. El lenguaje de consulta permite acceder a los datos, modificarlos y crearlos. Se divide en *lenguaje para la definición de datos* (acrónimo DDL[45]) y *lenguaje de gestión de datos*

[45] DDL es el acrónimo en inglés Data Definition Language, que es el que habitualmente se utiliza en la literatura de bases de datos.

(DML[46]). El DML sirve para formular las operaciones con los datos. Debe tener definidas, al menos, todas las operaciones del Álgebra Relacional. Las operaciones básicas del Álgebra Relacional son cinco: (i) unión, (ii) diferencia, (iii) producto cartesiano, (iv) selección y (vi) proyección. Existen otras cuatro operaciones que pueden formularse en función de las básicas: cociente, intersección, concatenación y semiconcatenación.

La *unión* de dos relaciones, que deben ser compatibles[47], es otra relación que contiene todas las tuplas de las dos. La *diferencia* de dos relaciones es otra relación que contiene las tuplas de la primera menos las tuplas que pertenezcan a la segunda. El *producto cartesiano* de k relaciones es otra relación formada por la concatenación de las k relaciones, de forma que su dimensión es la suma de las dimensiones de cada una de las k relaciones, y sus tuplas se forman concatenando ordenadamente cada tupla de la primera relación con el resto de las tuplas de las otras relaciones. Por ejemplo, el producto cartesiano de Alumnos × Matrículas es la relación Alumnos_Matrículas con atributos: DNI, nombre, Apellidos, Teléfono, DNI, Asignatura. Tendrá 12 tuplas :

> Alumnos × Matrículas = {(22222, Pepe, Jiménez, 91222222, 2222, Sintaxis), (22222, Pepe, Jiménez, 91222222, 3333, Sintaxis), (22222, Pepe, Jiménez, 91222222, 4444, Lingüística Informática), …., (4444, Concha, Pérez, 90000898, 4444, Lingüística Informática)}

La operación de *selección* (θ) selecciona tuplas o filas de una relación de acuerdo a una condición. Por ejemplo:

$$\theta_{DNI\ <4444}(\text{Alumnos}) = \{(2222, \text{Pepe, Jiménez, } 91222222), (3333, \text{María, Garcia, } 35889900)\}$$

Esta operación selecciona dos de las tres tuplas de la tabla Alumnos, las que tienen un valor de DNI inferior a 4444.

La operación de *proyección* (π) selecciona todos los valores de los atributos –columnas- que se especifican en la condición. Por ejemplo:

> $\pi_{DNI,teléfono}$(Alumnos) = {(2222, 91222222), (3333, 35889900), (4444, 90000898)}

[46] DML, Data Management Language.

[47] Dos relaciones son compatibles cuando tienen el mismo número de atributos y sus dominios son compatibles.

Todas estas operaciones se pueden combinar secuencialmente para resolver consultas más complejas. Por ejemplo, para obtener las asignaturas en las que está matriculado el alumno Pepe Jiménez tenemos que seleccionar primero en la tabla alumnos la tupla de Pepe Jiménez, luego la combinamos con un producto cartesiano con Matrículas, después elegimos aquella tupla del producto cartesiano que tiene el mismo valor en del DNI (Alumnos.DNI = Matrículas.DNI), y, finalmente, proyectamos sobre el atributo Asignaturas. La secuencia de operaciones y el resultado es el siguiente:

$$\pi_{Asignatura}(\theta_{Alumnos.DNI=Matriculas.DNI}(\theta_{nombre=\text{"Pepe"} \wedge apellido =\text{"Jiménez"}}(Alumnos) \times Matrículas)) = \{(Análisis del Discurso), (Sintaxis)\}$$

Los Sistemas Gestores de Bases de Datos tienen incorporado uno o más lenguajes de consulta para poder expresar las operaciones del Álgebra Relacional. El lenguaje de consulta de bases de datos relacionales más extendido es el SQL (ISO/IEC 9075-14, 2008). Es un estándar ISO/ANSI[48]- DDL y DML, de carácter declarativo, que incluye todas las operaciones del Álgebra Relacional, y, además, a partir de la versión ANSI SQL 2003, incluye consultas recursivas para el tratamiento de estructuras jerárquicas, funciones analíticas para el tratamiento de agregaciones y consultas sobre documentos marcados con el metalenguaje XML.

Las consultas en SQL se expresan con la sentencia básica SELECT:

```
SELECT Atributos
FROM Relaciones
WHERE condiciones;
```

La parte marcada con SELECT corresponde a una proyección, porque selecciona los atributos que se van a mostrar en la relación resultado. La parte FROM selecciona las relaciones que van a participar en la operación. Si son dos o más, se realiza el producto cartesiano. La parte WHERE corresponde a una selección según las condiciones que se escriban. Por ejemplo, la selección $\theta_{DNI <4444}(Alumnos)$, se expresa en SQL (el asterisco significa todos los atributos):

```
SELECT *
FROM Alumnos
WHERE DNI <44444;
```

La proyección $\pi_{DNI,teléfono}(Alumnos)$, se corresponde con la siguiente sentencia:

```
SELECT DNI, teléfono
FROM Alumnos;
```

[48] Está definido en el ISO/IEC 9075 y la última versión es de Julio de 2008 (SQL Language Reference, 2008).

Y no tiene clausula WHERE porque no hay selección. Finalmente el último ejemplo de composición secuencia de operaciones del Álgebra Relacional $\pi_{Asignatura}(\theta_{Alumnos.DNI=Matriculas.DNI}(\theta_{nombre=”Pepe” \wedge apellido =”Jiménez”}(Alumnos) \times Matrículas))$, se expresa en SQL:

```
SELECT Asignatura
FROM Alumnos, Matrículas
WHERE (Alumnos.DNI=Matrículas.DNI) AND (Alumnos.Nombre = “Pepe”) AND
        (Alumnos.Apellido =”Jiménez”);
```

El modelo relacional es un modelo diferente de los modelos revisados en las secciones anteriores. Es un modelo orientado al valor, con un lenguaje de carácter declarativo y con una estructura de organización de datos plana, basada en relaciones. Los modelos de datos jerárquico, de red y orientado a objetos identifican de forma única cada objeto con un identificador. Los lenguajes de consulta son, normalmente, de carácter procedimental y las estructuras de organización de datos son jerárquicas y reticulares. Desde el punto de vista de organización de la información léxica de los tesauros el modelo relacional, plano, no parece el más adecuado para implementar tesauros. Sin embargo, tiene dos ventajas importantes: (i) se dispone de un amplio abanico de SGBD, comerciales y de libre distribución, capaces de gestionar grandes volúmenes de datos de forma muy eficaz; (ii) se dispone de un sólido fundamento teórico para diseñar bases de datos muy optimizadas, con una redundancia mínima y un fuerte control de la consistencia de los datos.

Probablemente, por estas razones, sea el modelo de implementación de datos utilizado en la mayoría de los tesauros informatizados. Incluso, muchos de los actuales tesauros basados en lenguajes de marcado XML almacenan los datos en bases de datos relacionales y utilizan el formato XML para el intercambio de datos.

El diseño de la base de datos del tesauro no es una cuestión trivial, porque es necesario transformar las estructuras jerárquicas y las agregaciones del tesauro en estructuras planas. En el dominio de la medicina, Rada y Martin, (1987) proponen la construcción de tesauros aumentando o integrando tesauros ya existentes en una única base de datos relacional. Su propuesta se aplica a la integración de los tesauros de medicina MeSH[49], CMIT[50] y SNOMED[51]. El esquema de datos común es la unión de los esquemas particulares de cada tesauro, pero se añade un esquema de relación más que contiene

[49] http://www.nlm.nih.gov/mesh/
[50] http://www.nlm.nih.gov/mesh/presentations/NM2004_feb/man_machines/sld006.htm
[51] http://www.ihtsdo.org/snomed-ct/

cómo se relacionan los términos de los tres tesauros entre sí. Las estructuras jerárquicas de las relaciones más genera y más específico (TG/TE) se resuelven con dos relaciones:

Arbol(Código_Término_General, Código_Término_Específico)
Camino(Código_Termino_raiz, Código_Término_nivel_2,
Código_Término_nivel_3,…, Código_Término_nivel_9)

La primera almacena los pares término genérico y término específico, mientras que la segunda especifica los caminos desde la raíz hasta las hojas dentro de las jerarquías.
Las relaciones de sinonimia se resuelven dando un mismo código a todos los términos equivalentes:

Términos (Nombre_Término, Código_Termino)

Los resultados más destacables de esta aproximación son los siguientes: 1) la representación en el modelo relacional de las relaciones de jerarquía y sinonimia; 2) la integración de tesauros que, según apuntan los autores, se facilita con el uso del modelo relacional; 3) la eficiencia del modelo relacional, frente a otros modelos procedentes de la Inteligencia Artificial; 4) la necesidad de cierta intervención humana en la construcción o integración de los contenidos de los tesauros. Desde el punto de vista de escalabilidad y flexibilidad, esta aproximación propone un esquema de datos limitado y dependiente de la macroestructura de los tesauros que se están utilizando; por ejemplo, las jerarquías tienen un máximo de 9 niveles, ¿qué ocurre si se incluyen términos en un décimo nivel?. En el esquema de datos tampoco se incluyen las relaciones asociativas.
En Jones (1993), se propone un esquema de datos relacional más cercano al estándar de los tesauros, y, en consecuencia, más general, ya que incluye las relaciones TG/TE, equivalencia, asociativas, categorías:

Término (termino, cualificador, num_term, estatus, estructura, etc)
Palabras (palabra, frecuencia, raiz, sufijo)
Categoría (nivel, código_categoria, nombre_categoria)
Componentes (palabra, num_term)
Clases (código_categoria, núm_term)
Equivalencia (num_term_preferido, num_term_no_preferido, tipo)
Jerárquica (num_term_generico, num_term_especifico, tipo)
Asociativa (num_term, num_term, tipo)

En este esquema de datos se ha eliminado la relación ‘camino’ que no es general, y especifica todos los caminos posibles de términos desde la raíz hasta las hojas del árbol jerárquico TG/TE. También se han incorporado las relaciones asociativas. Sin embargo, a cambio, este esquema debe ser aumentado con procedimientos o programas que incluyan las transitividades de las relaciones jerárquicas y asociativas: si dos términos

están relacionados –con TG o TR- y el segundo término está relacionado con un tercero, el primero está relacionado con el tercero. Esto significa que para acceder a la información del tesauro, además de un SGBDs y un lenguaje de consulta general, se debe integrar una aplicación específica que gestione de forma completa toda la información de la base de datos.

Además, si el tesauro incluye otro tipo de relaciones semánticas específicas del dominio también es necesario crear estas relaciones –tablas- y añadirlas al esquema estándar. El resultado es una propuesta con un esquema de datos simple y estándar, pero con un sistema de gestión de la base de datos del tesauro complejo de crear, mantener, consular y, de nuevo, poco escalable.

Posteriormente, en Ballew, et al. (1999) se presenta un informe de la Universidad californiana de Berkeley, que examina las distintas alternativas para implementar las jerarquías de los tesauros con el modelo relacional, con el objetivo de conseguir una mayor flexibilidad en la representación y eficiencia en la consulta de la información. Utiliza el modelo de grafos como metamodelo para analizar las distintas soluciones que se han propuesto: 1) una relación con un atributo por cada nivel de la jerarquía; 2) una relación por cada nivel y otra relación que enlaza los niveles; 3) una relación en la que cada fila contiene un par término genérico, término específico y un tercer atributo con el camino desde la raíz hasta ese par de términos (tabla 5.3); finalmente, la última solución es 4) una relación para almacenar cada nodo de una jerarquía junto con el orden en que se visitarían los hijos izquierdo y derecho en un recorrido preorden (tabla 5.4). Esta última solución es la preferida por los autores del informe porque, con respecto a las otras soluciones, es más general, flexible y eficiente para almacenar y recuperar las relaciones jerárquicas del tesauro:

TG	TE	Camino
A	A	A
A	B	A,B
A	C	A,C
B	D	A,B,D
B	E	A,B,E

Tabla 5.3. Representación de la jerarquía con raíz A, hijo izquierdo B, hijo derecho C, hijo izquierdo de B, D, e hijo derecho de B, E.

Término	TE_izq	TE_dcho
A	1	10
B	2	7
C	8	9
D	3	4
E	5	6

Tabla 5.4. Representación de la jerarquía de la tabla 5.3 con la indicación del orden de visitación de los nodos (pre-orden): A, B, D, E, B, C

El resto de la base de datos relacional contiene los esquemas propuestos en Jones (1993). Además, incluye otra relación con la definición de los tipos de relaciones semánticas del tesauro. El esquema tiene un total de 14 relaciones. Efectivamente, es un esquema general y flexible para el almacenamiento y recuperación de la información del tesauro, pero es difícil de mantener por la complejidad de los cambios en las jerarquías. Si se inserta un término en medio de un camino jerárquico, hay que cambiar el orden de consulta de todos los nodos de esa jerarquía, lo cual es costoso. Al igual que en la solución de Jones, no está prevista la representación de estructuras jerárquicas de inclusión de categorías y subcategorías, lo cual resta algo de generalidad al esquema.

Otras aproximaciones con este modelo relacional incorporan la posibilidad de gestión colaborativa de la información, de documentación y ciclo de vida de los términos, simplemente añadiendo relaciones extra para incorporar esta información (Pastor y Martínez, 2003), pero esencialmente son soluciones que repiten los esquemas presentados.

El modelo relacional se ha utilizado, por razones de eficiencia y accesibilidad de los SGBDs, para la implementación del modelo conceptual propuesto en este trabajo de investigación.

5.3.2. Modelos basados en Lenguajes de marcado XML

Los lenguajes de marcado tienen el propósito de describir explícitamente el modelo de organización del contenido de un recurso web, documento o conjunto de datos en formato electrónico (Leech, 2005). Los recursos con contenido textual, como los tesauros, se pueden enriquecer con marcas añadidas al texto para hacer explícita la estructura o la semántica de dicho texto con respecto de un modelo de datos. Este modelo de datos puede, además, estar definido formalmente en una gramática del

lenguaje de marcas DTD o Schema. El uso de marcas para describir el contenido textual de los tesauros no es una novedad, porque las normas estándares de construcción de tesauros, revisados anteriormente, también definen un conjunto de marcas estándares -TG y TE, USE y USE PARA, TR- para describir las relaciones entre los términos.

Actualmente, en el entorno de la Web, los documentos están marcados con lenguajes definidos mediante el metalenguaje estándar eXtensible Markup Language (XML) (XML, 2008). XML surge en el año 1998 con el fin de simplificar el metalenguaje SGML, que era más complejo aunque más completo y potente, para facilitar y promover el uso de los lenguajes de marcado en los documentos y datos que se almacenan y trasmiten por Internet. Su uso favorece la aparición de aplicaciones software de carácter general para procesar estas marcas XML y así poder compartir, transmitir, interpretar y manipular cualquier tipo de información que esté previamente marcada con XML. Los ejemplos más conocidos de aplicaciones XML son los navegadores. Las aplicaciones más avanzadas son los "agentes inteligentes", capaces de interpretar y actuar, adaptándose a diferentes situaciones según el contexto, y capaces de comunicarse enviando y recibiendo mensajes. XML ha sido definido, es revisado y actualizado por el consorcio W3C, responsable de todos los estándares de la Web.

Un lenguaje XML es un modelo para organizar y describir información[52] (TEI-P5, 2007)[53] Este modelo define la estructura de la información, mediante un conjunto de reglas independientes del contexto y, opcionalmente, con una lista de atributos y valores de los componentes de la estructura. En la figura 5.19 se muestra la gramática XML que define la estructura del tesauro Agrícola de la National Agricultural Library. Como se puede comprobar, el tesauro, que es el elemento raíz, THESAURUS, consta de uno o más CONCEPT[54]. A su vez cada CONCEPT está formado por un DESCRIPTOR o bien por un NON-DESCRIPTOR seguidos, opcionalmente, de uno o más elementos del tipo SC, DF, SN, UP, USO, BT, NT, RT,…, FLG. Finalmente, cada uno de estos últimos elementos contiene texto o texto marcado (PCDATA).

[52] Uno de los lenguajes XML más conocidos es el XHTML, que es una versión definida en XML del antiguo HTML, definido con SGML.

[53] Disponible en: http://www.tei-c.org/Guidelines/P5/

[54] El '+' simboliza una o más apariciones del elemento que le precede, por su parte el símbolo '*' significa cero, uno o más apariciones.

```
<!DOCTYPE THESAURUS [
<!ELEMENT THESAURUS (CONCEPT+)>
<!ELEMENT CONCEPT ( (DESCRIPTOR|NON-
DESCRIPTOR),SC*,DF*,SN*,UP*,USO*,BT*,NT*,RT*,EN*,TNR*,STA*,IN
P*,APP*,UPD*,NVD*,FLG*)>
<!ELEMENT DESCRIPTOR (#PCDATA)>
<!ELEMENT NON-DESCRIPTOR (#PCDATA)>
<!ELEMENT SC (#PCDATA)>
<!ELEMENT DF (#PCDATA)>
<!ELEMENT SN (#PCDATA)>
<!ELEMENT UP (#PCDATA)>
<!ELEMENT USO (#PCDATA)>
<!ELEMENT BT (#PCDATA)>
<!ELEMENT NT (#PCDATA)>
<!ELEMENT RT (#PCDATA)>
<!ELEMENT EN (#PCDATA)>
<!ELEMENT TNR (#PCDATA)>
<!ELEMENT STA (#PCDATA)>
<!ELEMENT INP (#PCDATA)>
<!ELEMENT APP (#PCDATA)>
<!ELEMENT UPD (#PCDATA)>
<!ELEMENT NVD (#PCDATA)>
<!ELEMENT FLG (#PCDATA)>
```

Figura 5.19. Gramática del lenguaje de marcado XML del tesauro Agrícola de la Nacional Agricultural Library (edición 2008).[55]

Como se puede apreciar en el ejemplo, los lenguajes XML son modelos de datos de naturaleza jerárquica, puesto que utiliza reglas independientes del contexto para definir la estructura de la información. Siempre debe existir un elemento inicial o raíz a partir del cual se definen, de forma arborescente, todos los demás.

Una vez definida la gramática o modelo de un lenguaje de marcas XML, éste se puede utilizar en cualquier documento cuya información sea compatible con el modelo. En la figura 5.20, se muestra el uso del lenguaje de marcado definido en la figura anterior. En el proceso de marcado se identifica cada parte estructural del documento y se marca con un par de marcas o etiquetas, una de apertura y otra de cierre. Se puede comprobar que todo el tesauro está delimitado por la etiqueta de apertura <THESAURUS> y la de cierre </THESAURUS>. A su vez, cada entrada se marca por las etiquetas <CONCEPT> y </CONCEPT>. La primera entrada de la figura está formada por un término descriptor, <DESCRIPTOR>: 'abejas carpinteras', un término genérico, <BT>: 'plagas de insectos', un término relacionado, <RT>: 'Xylocopa', la traducción del

[55] Disponible en : http://agclass.nal.usda.gov/agt.shtml (en inglés), o
En: http://agclass.nal.usda.gov/agt_es.shtml (en español)

descriptor al inglés, <EN>: ‘carpenter bees’ y, finalmente, un número de entrada, <TNR>: ‘22622’.

```
<THESAURUS>
      <CONCEPT>
          <DESCRIPTOR>abejas carpinteras</DESCRIPTOR>
           <BT>plagas de insectos</BT>
           <RT>Xylocopa</RT>
           <EN>carpenter bees</EN>
           <TNR>22622</TNR>
      </CONCEPT>

      <CONCEPT>
          <NON-DESCRIPTOR>abejas mieliferas</NON-DESCRIPTOR>
           <USO>abejas mielíferas</USO>
           <TNR>98496</TNR>
      </CONCEPT>

      <CONCEPT>
          <DESCRIPTOR>abejas mielíferas</DESCRIPTOR>
           <UP>abejas mieliferas</UP>
           <BT>insectos</BT>
           <NT>abejas africanizadas</NT>
           <NT>abejas obreras</NT>
           <NT>abejas reinas</NT>
           <NT>abejas sin aguijón</NT>
           <NT>panales de zánganos</NT>
           <RT>apicultura</RT>
           <RT>Apidae</RT>
           <RT>Apis mellifera</RT>
           <RT>criadero de abejas mielíferas</RT>
           <RT>insectos benéficos</RT>
           <RT>miel y productos fabricados por abejas</RT>
           <RT>Nasonovia feromona</RT>
           <EN>honey bees</EN>
           <TNR>4918</TNR>
      </CONCEPT>
……
</THESAURUS>
```

Figura 5.20. Aplicación del lenguaje de marcado definido en la figura 5.19 a la información del tesauro Agrícola de la NAL.

La ventaja de XML es la flexibilidad para construir cualquier modelo o gramática con este metalenguaje común. Pero, al mismo tiempo, la flexibilidad puede ser una desventaja, porque se corre el peligro de que la Web se transforme en una ‘torre de Babel’ de lenguajes XML en la que ya no sea posible compartir la información ni garantizar la interoperabilidad de las aplicaciones. En consecuencia, es importante tener en cuenta que en la creación y selección de un modelo XML la precisión y la

exhaustividad son criterios que se contraponen al nivel de generalización y grado de difusión del modelo: cuánto más completo y preciso sea un modelo mejor se ajustará al dominio de información pero, a la vez, menos general será y menos difundido estará. Actualmente la práctica más frecuente es el uso de los *perfiles de aplicación*. Un perfil de aplicación es un modelo estándar general que se ha particularizado en un dominio, siguiendo unas normas estándar preestablecidas.

La aceptación y difusión de los modelos generales y estándares dependen fundamentalmente, de dos factores: (i) si son sencillos de utilizar y (ii) si se dispone de software para su explotación; los modelos XML generales pueden ser complicados de entender y aplicar, y en algunos casos no existen aplicaciones porque las compañías de software intentan condicionar el mercado desarrollando aplicaciones para sus propios modelos, de forma que una vez construido el sistema no es posible exportarlo y utilizarlo en otras aplicaciones. Entre las estrategias que se adoptan para resolver este problema está el promover, desde los consorcios internacionales de los estándares, la creación de aplicaciones y herramientas de código abierto o de libre distribución que faciliten tanto el uso de los modelos generales como la explotación de la información marcada. También en esta línea, se invita a las compañías de software a formar parte de estos consorcios que desarrollan y difunden los modelos estándar XML.

La nueva generación de aplicaciones Web se construye ya para la Web Semántica. El objetivo de la Web Semántica es organizar y explotar colaborativa e inteligentemente la información. Las aplicaciones de la Web Semántica utilizan los tesauros y las ontologías para dar significado a los documentos y recursos de la Web y poder buscarlos y explotarlos a partir de estas descripciones semánticas. Sin embargo, esto sólo es posible si el contenido de los tesauros y los modelos XML para definir este contenido son estándares, generales, consensuados y difundidos. La estandarización del contenido de los tesauros y ontologías implica: i) consensuar una única forma de organizar y expresar la información y el conocimiento, lo cual es realmente difícil incluso en dominios de especialidad, o ii) encontrar mecanismos de integración y correspondencia entre tesauros (Stuckenschmidt y Harmelen, 2005). Para resolver la segunda cuestión, la integración de los contenidos de tesauros diferentes, se proponen actualmente varias iniciativas[56], entre las que destacan utilizar: 1) modelos basados en RDF para la Web Semántica, que son promovidos por el World Wide Web Consortium

[56] Una buena página para consultar tesauros XML en diferentes formatos es http://www.w3c.rl.ac.uk/SWAD/thes_links.htm

(w3c[57]); y 2) modelos procedentes del *e-learning* para la definición de los vocabularios de los metadatos, que son promovidos por los organismos internacionales para la difusión y elaboración de propuestas de estándares en e-learning –CEN[58], IMS Global Learning Consortium[59].

5.3.2.1. Modelos basados en el Resource Description Framework (RDF)

El Marco de Descripción de recursos RDF

El modelo RDF es, en realidad, un metamodelo para crear modelos que describan semánticamente los objetos de la Web (RDF, 2004). Este metamodelo se basa en dos supuestos: 1) los objetos del dominio son recursos y se identifican con un URI (Uniform Resource Identification), que es un identificador único en la Web; y 2) los objetos se describen mediante un conjunto de pares atributo y valor. Los valores de las propiedades pueden ser, a su vez, objetos.

El conjunto de objetos, propiedades y valores puede representarse visualmente con un grafo (Figura 5.21). Los nodos son objetos o valores de las propiedades. Los arcos representan las propiedades, con una etiqueta que es el identificador de la propiedad. La figura 5.21 muestra una representación gráfica del modelo de organización RDF, utilizando, como ejemplo, el término *Ecosystems* del tesauro CERES[60]. El término *Ecosystems* tiene como forma ortográfica (Label) "Ecosystems" y como Término Genérico (BT), el Descriptor0101 con forma ortográfica (Label) "Biosphere". Ambos términos se consideran recursos Web de tipo Descriptor con identificadores 010101 y 0101 respectivamente.

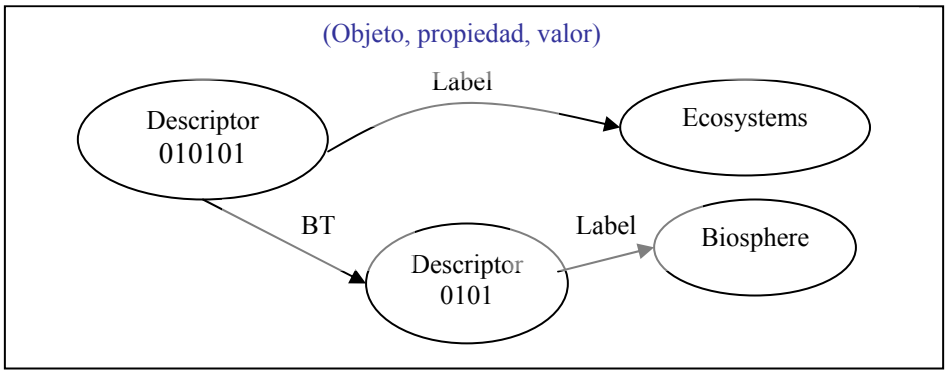


Figura 5.21. Grafo RDF basado en la estructura nodo:objeto, arco: propiedad, nodo: valor. Fuente de datos CeresRDF

[57] http://www.w3.org/
[58] European Committee for Standarization: http://www.cen.eu/
[59] http://www.imsglobal.org/
[60] http://ceres.ca.gov/thesaurus/Overview.html

Los grafos RDF se representan habitualmente con lenguajes XML. A esta traducción se le llama *serialización* del grafo RDF en RDF/XML. En la figura 5.22 se presenta una posible serialización[61] RDF/XML del grafo de la figura 5.21.

```
1) <?xml version="1.0"?>
2) <?xml:namespace ns='http://www.w3.org/TR/WD-rdf-syntax/'
prefix='RDF' ?>
3) <?xml:namespace ns='http://www.w3.org/TR/WD-rdf-schema/'
prefix='RDFS' ?>
4) <?xml:namespace ns='http://ceres.ca.gov/thesaurus/' prefix='Z19' ?>
5) <RDF:RDF>
6) <Descriptor RDF:id="010101">
  <Label>Ecosystems</Label>
  <BT>
      <Label>Biosphere</Label>
      <Descriptor RDF:resource="0101"/>
  </BT>
 </Descriptor>
 </RDF:RDF>
```

Figura 5.22. Versión RDF/XML de la figura 5.31[62].

La interpretación de esta versión RDF/XML es la siguiente: la primera línea declara que se trata de una representación XML; la segunda línea declara el uso del metamodelo RDF. Esta declaración se hace utilizando el mecanismo de XML para definir modelos denominados "espacio de nombres"[63]; la tercera y cuarta líneas declaran, respectivamente, el uso del modelo RDF-Schema, que presentaremos en la siguiente sección, y del modelo del tesauro CERES; en ambos casos se indica la URI del documento textual donde está definido el modelo; la quinta línea indica el comienzo del contenido del documento –el tesauro en este caso. La etiqueta de esta quinta línea es el elemento raíz del documento RDF/XML <RDF:RDF>; el resto de las marcas de esta figura forman parte del modelo particular del tesauro CERES, declarado en la cuarta línea, y que ha sido definido por los autores del tesauro.

Una aplicación del modelo de implementación RDF para definir un esquema conceptual se muestra en la figura 5.23. Este esquema RDF es la traducción del esquema Orientado a Objetos de la figura 5.18, creado para describir la estructura estándar de los tesauros:

[61] Existen varias serializaciones posibles para este grafo, pero no todas son equivalentes computacionalmente ni respecto a la legibilidad para humanos.
[62] Los números de línea no forman parte del documento, se han incluido sólo para su descripción.
[63] El espacio de nombres es la traducción literal del término XML Namespace. Se refiere al documento o documentos de texto que describen un modelo. En estos documentos deben ser accesibles via Web y contienen la definición de todos los elementos estructurales y sus propiedades correspondientes a ese modelo.

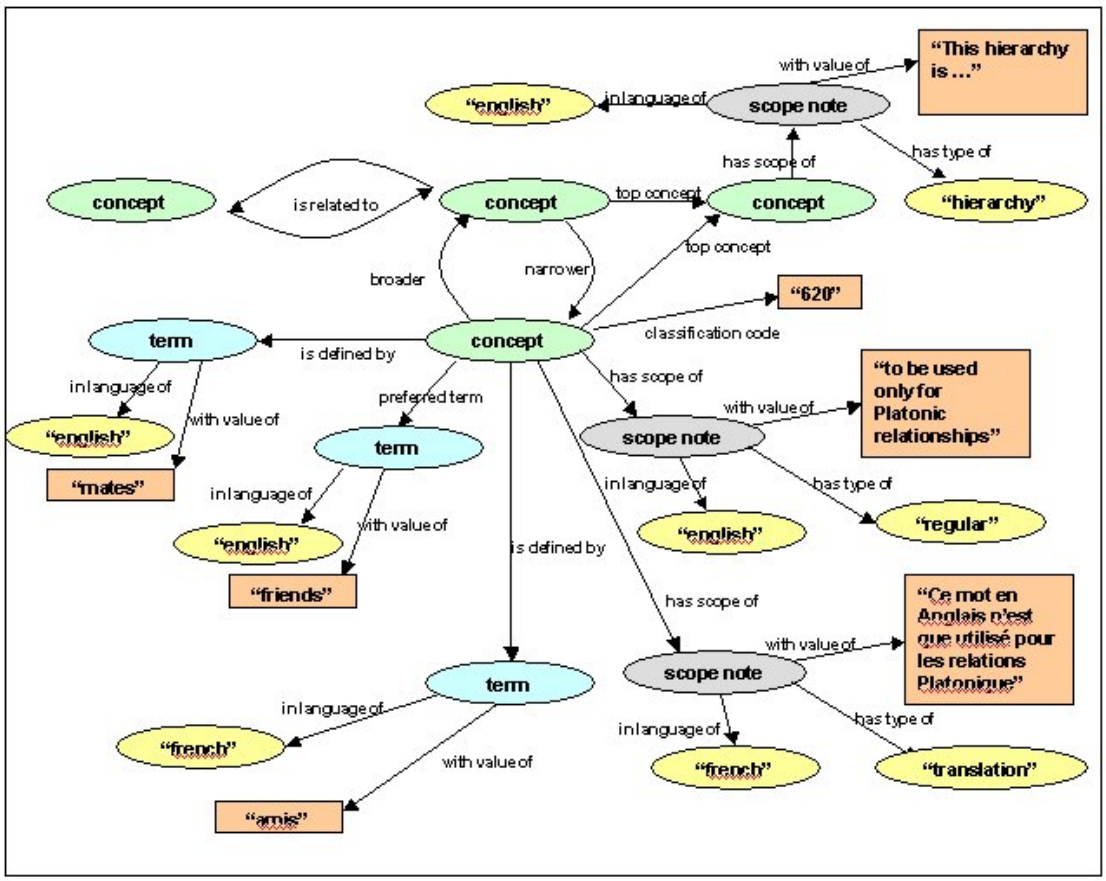


Figura 5.23. Esquema RDF para la construcción de tesauros según el esquema conceptual de la figura 5.18[64]

Con este esquema se ha construido el Tesauro de Ciencias Sociales ELSST[65]. En la figura 5.24 se muestra, como ejemplo, la entrada 'economics' utilizando una serialización XML.

[64] Fuente: http://www.w3c.rl.ac.uk/pasttalks/slidemaker/XML_UK_SW_Thes/slide14-0.html
[65] European Language Social Science Thesaurus (Balkan, et. al, 2002).

```
<rdf:RDF xml:lang="en">

<rdf:Description rdf:ID="EN-N">
<rdf:type rdf:resource="http://www.limber.rl.ac.uk/External/thesaurus-iso.rdf#TopConcept"/>
<thes:ClassificationCode>N</thes:ClassificationCode>

<thes:PreferredTerm>
<rdf:Description>
<rdf:type rdf:resource="http://www-limber.rl.ac.uk/External/thesaurus-iso.rdf#Term"/>
<thes:inLanguageOf rdf:resource="http://www.limber.rl.ac.uk/External/ISO639.rdf#en"/>
<rdf:value>ECONOMICS</rdf:value>
</rdf:Description>
</thes:PreferredTerm>

<thes:UsedFor>
<rdf:Description>
<rdf:type rdf:resource="http://www.limber.rl.ac.uk/External/thesaurus-iso.rdf#Term"/>
<thes:inLanguageOf rdf:resource="http://www.limber.rl.ac.uk/External/ISO639.rdf#en"/>
<rdf:value>POLITICAL ECONOMY</rdf:value>
</rdf:Description>
</thes:UsedFor>
<thes:NarrowerConcept rdf:ID="N50-59"/>
<thes:NarrowerConcept rdf:ID="N98-20"/>
<thes:NarrowerConcept rdf:ID="N35-50"/>
<thes:NarrowerConcept rdf:ID="Q21"/>
<thes:NarrowerConcept rdf:ID="N90-10-10"/>
.....
<thes:NarrowerConcept rdf:ID="R89"/>
<thes:RelatedConcept rdf:ID="Q45-99"/>

<thes:ExactEquivalent>
<rdf:Description rdf:ID="FR-N">
<thes:inLanguageOf rdf:resource="http://www.limber.rl.ac.uk/External/ISO639-rdf#fr"/>
</rdf:Description>
</thes:ExactEquivalent>

<thes:ExactEquivalent>
<rdf:Description rdf:ID="DE-N">
<thes:inLanguageOf rdf:resource="http://www.limber.rl.ac.uk/External/ISO639-rdf#de"/>
</rdf:Description>
</thes:ExactEquivalent>

<thes:ExactEquivalent>
<rdf:Description rdf:ID="SP-N">
<thes:inLanguageOf rdf:resource="http://www.limber.rl.ac.uk/External/ISO639-rdf#sp"/>
</rdf:Description>
</thes:ExactEquivalent>
</rdf:Description>
</rdf:RDF>
```

Figura 5.24. Entrada Economics del tesauro ELSST de Ciencias Sociales en formato RDF/XML[66]

[66] Fuente: http://www.w3c.rl.ac.uk/pasttalks/slidemaker/XML_UK_SW_Thes/ELSST_economics.xml

En resumen, la información en la Web puede organizarse mediante un metamodelo basado en grafos denominado RDF (figura 5.25). El tipo de nodos y arcos del grafo forman el modelo de datos del dominio de información. Los modelos de datos, tanto si son estándar como propietario, se implementan con el lenguaje RDFS. En la sección siguiente, se presenta, con un ejemplo tomado de Cross et al. (2000), una breve revisión de RDFS y su uso.

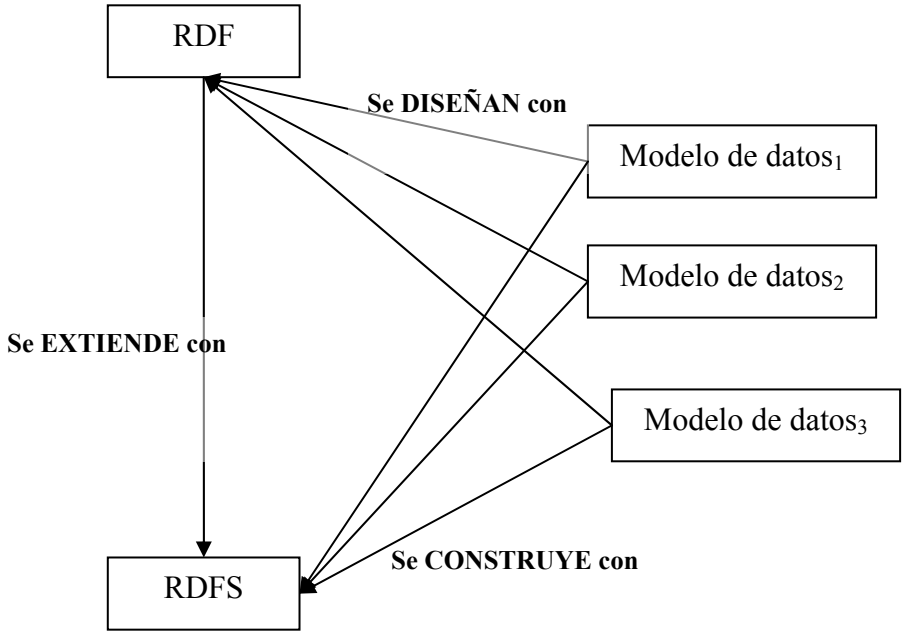


Figura 5.25. Relación conceptual entre RDF, RDFS y los modelos de datos

Resource Description Framework Schema (RDFS)

El esquema RDF, RDFS, es un lenguaje de propósito general basado en RDF para la descripción de los modelos de datos diseñados con RDF (RDFS, 2004). Puede considerarse un lenguaje del tipo orientado a objetos, pero restringido a la definición de las "estructuras de datos" ya que no incluye definición de la funcionalidad.

El propósito del RDFS es definir y dar semántica a los grafos RDF. RDF permite dibujar un esquema gráfico de recursos, propiedades y relaciones, mientras que RDFS es un lenguaje para definir el esquema RDF y, además, permite la organizar los recursos en *clases* o tipos cuando comparten una serie de propiedades en común (Figura 5.26).

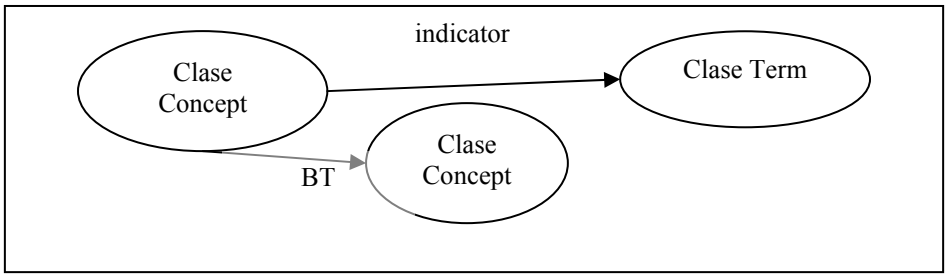


Figura 5.26. Modelo de tesauro: definición de clases, propiedades y relaciones[67]

Una *clase* se define en RDFS con un *esquema de clase*, que es una plantilla identificada con un único URI y caracterizada por un conjunto de propiedades. Los tipos de relaciones entre recursos se definen también mediante propiedades (*rdfs:propiert)*. En RDFS existen dos relaciones predefinidas: (i) la pertenencia de un recurso a una clase, que también se llama instancia, se explicita con el atributo *rdfs:type*[68] y (ii) la inclusión de una clase en otra, mediante la relación r*dfs:subClassOf.* En la figura 5.27 las nuevas clases *Concept* y *Term* se definen como subclases de la clase inicial *Resource*.

[67] Adaptado de Cross et al. (2000).
[68] Un recurso que pertenece a una clase se denomina "instancia", igual que en POO.

```
<rdfs:Class rdf:ID="Concept">
   <rdfs:subClassOf
     rdf:resource="http://www.w3.org/TR/1999/PR-rdf-schema-19990303#Resource"/>
   </rdfs:Class>

    <rdfs:Class rdf:ID="Term">
      <rdfs:comment>
          Instances of this class represent the written forms of Concepts. The string is
          given by the rdf:value of Term.
      </rdfs:comment>
      <rdfs:subClassOf
        rdf:resource="http://www.w3.org/TR/1999/PR-rdf-schema-
19990303#Resource"/>
   </rdfs:Class>

   <rdf:Property ID="broaderConcept">
     <rdfs:comment>
          This schema does not define a property 'narrowerConcept', but applications can
          assume the existence of a property narrowerConcept such that if:
       {broaderConcept,ConceptA,ConceptB}, then
       {narrowerConcept,ConceptB,ConceptA} is true.
     </rdfs:comment>
     <rdfs:domain rdf:resource="#Concept"/>
     <rdfs:range rdf:resource="#Concept"/>
   </rdf:Property>

   <rdf:Property ID="indicator">
      <rdfs:comment>
          A mandatory property of a Concept whose value is the Term instance representing
          a written form of the Concept. A Concept may have as an indicator more than one
          Term. A Term may only be an indicator of one Concept.
      </rdfs:comment>
      <rdfs:domain rdf:resource="#Concept"/>
      <rdfs:range rdf:resource="#Term"/>
   </rdf:Property>
```

Figura 5.27. Definición RDFS del modelo de la figura 5.36 con XML.

El conjunto de recursos que conforman un esquema de clase forman la *extensión de la clase* (*class extension*) [69]. Cada recurso perteneciente a una clase es una instancia. En la figura 5.28 se observa la definición de una instancia de la clase Concept, el concepto llamado CID_6. Este concepto se define utilizando el modelo RDF de la figura 5.26 y el lenguaje RDFS de la figura 5.27. En negrita se destacan las relaciones y propiedades del concepto. Se utiliza el atributo *type* para indicar la pertenencia a la clase Concept; el atributo *indicador* relaciona el concepto con el término TID_3; el atributo *broaderConcept* indica el concepto más general (CID_8); y, finalmente, el atributo *relatedConcept* indica la relación asociativa TR con el concepto CID_15.

[69] Un conjunto de recursos puede ser una extensión de dos clases diferentes: los alumnos de la *clase EstudiantesLinguisticaComputacional* pueden ser exactamente los de la clase *EstudiantesEjemplares.*

```
<web:Description about="http://sosig.ac.uk/hasset/concepts/CID_6">

  <web:type resource="http://snowball.ilrt.bris.ac.uk/~pldab/rdf-
dot/Thes/Thes.xrdf#Concept"/>

  <rdfs:isDefinedBy web:resource="http://sosig.ac.uk/hasset/concepts/"/>

  <thes:indicator web:resource="http://sosig.ac.uk/hasset/terms/TID_3"/>

  <thes:broaderConcept>

    <web:Description about="http://sosig.ac.uk/hasset/concepts/CID_8">

      <thes:indicator web:resource="http://sosig.ac.uk/hasset/terms/TID_15"/>

      <thes:conceptCode>769</thes:conceptCode>

    </web:Description>

  </thes:broaderConcept>

  <thes:relatedConcept
web:resource="http://sosig.ac.uk/hasset/concepts/CID_15"/>

</web:Description>
```

Figura 5.28. Especificación del concepto CID_6 con el modelo RDF de la figura 5.25.

El modelo Ontology Web Language (OWL)

El *Lenguaje Web de Ontologías (OWL)* es un modelo estándar basado en RDF que amplía la semántica básica de RDFS con una semántica formal. Constituye una capa de abstracción mayor sobre RDFS cuyo objetivo es representar y procesar, incluyendo inferencias, el conocimiento en la Web. *OWL* consta de tres sublenguajes de expresividad creciente: OWL Lite, OWL DL, y OWL Full (Bechhofer, et. al, 2004). La relación sintáctica y semántica entre RDFS y los sublenguajes de OWL es de inclusión respecto del poder expresivo:

$$\text{RDFS} \subset \text{OWL Lite} \subset \text{OWL DL} \subset \text{OWL Full}$$

Cualquier representación RDFS es una representación OWL, pero las representaciones OWL tienen mayor poder expresivo que RDFS. OWL permite expresar relaciones entre clases, restricciones como la cardinalidad de las relaciones y operaciones booleanas como la igualdad/desigualdad entre clases o instancias, operaciones de conjuntos sobre las clase como el complemento o la unión.

El uso de OWL para la construcción de tesauros no es habitual, aunque el tesauro puede considerarse una ontología terminológica (Matthews et al., 2003). Se utiliza OWL para representar el contenido de los tesauros cuando se trata de 1) integrar en un mismo sistema varios tesauros del mismo dominio de conocimiento, o 2) cuando se utiliza el contenido de uno o varios tesauros como base para la construcción de una ontología. Sin embargo, estos casos no deben considerarse como aplicaciones de OWL a la construcción de tesauros sino una conversión o migración de tesauros a ontologías simples. Un ejemplo reciente es el Servidor de Conceptos AGROVOC (CS)[70], un sistema multilingüe en línea para la búsqueda de conceptos relacionados con la agricultura, que utiliza el contenido del tesauro AGROVOC, previamente transformado en una ontología (Sini, et. al, 2008).

El modelo Simple Knowledge Organisation (SKOS-Core)

El lenguaje *Simple Knowledge Organisation for the Web (SKOS-Core)* es una adaptación de OWL al estándar ISO 2788-1986 de tesauros (actualmente ANSI/NISO Z39.19, 2005), pero su aplicación se ha generalizado a la representación de cualquier *esquema de conceptos*[71] (Miles et al., 2005). SKOS representa el tesauro como una ontología compatible con el modelo tradicional de tesauros ANSI/NISO Z39.19.

Un tesauro se construye con SKOS-Core a partir del elemento Concepto. Este elemento está relacionado con uno o más términos preferidos y no preferidos. Los términos no están relacionados directamente, sino a través de los conceptos. Las relaciones semánticas entre los términos de un tesauro se trasladan a los conceptos asociados a dichos términos, por lo tanto son relaciones entre conceptos. La tabla 5.5 muestra la relación entre los elementos del modelo SKOS-Core y los elementos básicos del modelo estándar de construcción a los tesauros. La figura 5.30 muestra un ejemplo de aplicación.

ANSI/NISO Z39.19	SKOS-Core
Tesauro	Esquema de conceptos (skos:ConceptSchema)
	Concepto con un único URI (skos:Concept)
Término	
Término preferido (USE FOR)	skos:prefLabel
Término no preferido (USE)	skos: altLabel
Categoría	Esquema de Conceptos ó colecciones (skos:collection)
Pertenencia (término a categoría	skos:inScheme ó skos:member

[70] http://naist.cpe.ku.ac.th/agrovoc/

[71] Se entiende por esquemas de conceptos (concept schemas) un conjunto de conceptos con relaciones semánticas.

o tesauro)	
BT/NT (término génerico/término específico	skos:broader / skos:narrower (concepto genérico / concepto específico)
RT (término relacionado)	skos:related (concepto relacionado
SN (nota de ámbito)	skos:scopeNote

Tabla 5.5. Relación básica entre los modelos ANSI/NISO Z39.19 y SKOS-Core

```
<rdf:RDF>
<skos:ConceptScheme rdf:nodeID="tematres">
<skos:Concept rdf:nodeID="tema1258">
        <skos:prefLabel>enseñanza secundaria</skos:prefLabel>
        <skos:altLabel>bachillerato</skos:altLabel>
        <skos:inScheme rdf:nodeID="tematres"/>
        <skos:broader rdf:nodeID="tema3114"/>
        <skos:narrower rdf:nodeID="tema3039"/>
        <skos:narrower rdf:nodeID="tema3040"/>
        <skos:subjectIndicator
        rdf:resource="http://redined.r020.com.ar/es/index.php?tema=1258"/>
</skos:Concept>
</rdf:RDF>
```

Figura 5.30. Definición SKOS-Core del término enseñanza secundaria[72]

El modelo SKOS-Core incluye otros elementos, relaciones y tipos de datos que permite construir estructuras conceptuales de complejidad mayor que los tesauros (Miles y Bechhofer, 2008). Además, permite combinar otros modelos o lenguajes estándares, como el modelo Dublin Core, para aumentar su capacidad expresiva[73] (figura 5.31).

Consideraciones finales sobre los modelos basados en RDF

Una aproximación posible para la construcción de tesauros es el uso del metamodelo RDF para estructurar su contenido en grafos de elementos relacionados con otros elementos o con los valores de sus propiedades. Los elementos del tesauro tienen un identificador único Web (URI) y las relaciones toman nombres según el modelo RDF que se aplique. Los modelos RDF aplicados en la construcción de tesauros son propietarios o estándares. Entre los estándares destacamos el RDFS, que amplía

[72] Tomado de la fuente: Tesauro Europeo de la Educación, disponible en http://redined.r020.com.ar/es/

[73] La figura muestra la descripción de un recurso, el propio Tesauro Europeo de la Educación, utilizando metadatos y el contenido SKOS-Core del tesauro (en los capítulos 2 y 3 se revisa el uso de metadatos y vocabularios para la descripción de recursos).

semánticamente las estructuras RDF para 1) agruparlas en jerarquías de clases y en instancias cuando comparten relaciones y propiedades, y 2) definir restricciones. Los modelos OWL y SKOS-Core están basados en RDF y RDFS y tienen como objetivo la representación unificada de los vocabularios de la Web como ontologías. Son una apuesta importante del w3c en la evolución de la Web de la información hacia la Web semántica que precisa modelos que den significado a los recursos. Las ontologías y tesauros OWL y SKOS-Core son uno de los pilares de la Web semántica

```
<rdf:RDF>
<skos:ConceptScheme rdf:nodeID="tematres">
        <dc:title>Tesauro Europeo de la Educación</dc:title>
        <dc:creator>Comisión de las Comunidades Europeas; Councl of Europe;
        REDINED
        </dc:creator>
        <dc:subject>EDUCACION; ENSEÑANZA; TESAUROS DE
        EDUCACION</dc:subject>
        <dc:description>
         Tesauro de educación en base al procesamiento automático de la versión en
        lengua española del 2003.
        </dc:description>
        <dc:publisher>Comisión de las Comunidades Europeas; Councl of Europe;
        REDINED</dc:publisher>
        <dc:date>2003-12-31</dc:date>
        <dc:language>es</dc:language>
<skos:Concept rdf:nodeID="tema1258">
        <skos:prefLabel>enseñanza secundaria</skos:prefLabel>
        <skos:altLabel>bachillerato</skos:altLabel>
        <skos:hiddenLabel>bachillerato</skos:hiddenLabel>
        <skos:inScheme rdf:nodeID="tematres"/>
        <skos:broader rdf:nodeID="tema3114"/>
        <skos:narrower rdf:nodeID="tema3039"/>
        <skos:narrower rdf:nodeID="tema3040"/>
        <dct:created>2007-02-22 02:29:38</dct:created>
</skos:Concept>
</rdf:RDF>
```

Figura 5.31. Combinación de SKOS-Core y metadatos Dublín Core[74]

[74] Adaptado de la fuente: Tesauro Europeo de la Educación, disponible en http://redined.r020.com.ar/es/

Son modelos basados en grafos que heredan la versatilidad y poder expresivo de estas estructuras matemáticas. Sin embargo, presentan actualmente dos importantes inconvenientes: 1) no se dispone de suficientes herramientas de carácter general para el acceso y manipulación de las redes RDF, y 2) las herramientas actuales son ineficientes cuando se trata de procesar los grafos RDF de los tesauros reales porque el tamaño y complejidad de estas redes para un número de elementos relativamente pequeño (100) hace prohibitivo su procesamiento[75] (Miles, 2003).

5.3.2.2. Modelos procedentes del *e-learning*

IMS Vocabulary Definition EXchange (IMS VDEX)

La especificación *Definición de Intercambio de Vocabularios IMS* ó *IMS Vocabulary Definition EXchange* es una especificación del IMS Global Learning Consortium aprobada en Febrero de 2004[76]. Es una gramática para la representación e intercambio de los vocabularios (listas, taxonomías y tesauros) utilizados en las aplicaciones *e-learning* (IMS VDEX, 2004). Su objetivo es proporcionar un modelo y lenguaje XML común que fomente el intercambio y reusabilidad de los vocabularios de referencia utilizados en las propiedades de los esquemas de metadatos educativos –IEEE LOM, IMS Meta-Data, IMS Learner Information Package, ADL SCORM- que clasifican y/o definen conceptualmente los recursos didácticos. El uso de VDEX para la construcción de vocabularios y de los esquemas de metadatos educativos estándares garantiza la interoperabilidad entre repositorios de recursos didácticos digitalizados u objetos de aprendizaje.

El modelo VDEX organiza el contenido de los tesauros únicamente con los dos elementos términos y relaciones (figura 5.32). Es, en consecuencia, un modelo basado en términos. Además del modelo de contenido del tesauro, VDEX incluye un conjunto de propiedades –como *vocabName*- y metadatos para documentar el tesauro de forma global.

[75] En un experimento llevado a cabo para consultar la información en un grafo RDF con tan sólo 100 elementos (26118 sentencias) el tiempo de búsqueda superaba las 24 horas.
[76] http://www.imsglobal.org/vdex/

```
<?xml version="1.0" encoding="UTF-8" ?>
- <vdex orderSignificant="false" profileType="thesaurus" xsi:schemal
    xmlns="http://www.imsglobal.org/xsd/imsvdex_v1p0" xmlns
  + <vocabName>
  + <term>
  + <term>
  + <term>
  + <relationship>
  + <relationship>
  </vdex>
```

Figura 5.32. Estructura básica de un tesauro VDEX

La colección de términos puede estar ordenada o no. En el modelo IMS VDEX (figura 5.33) cada término es una estructura compleja formada por: 1) un conjunto de propiedades y valores que describen al término, incluso mediante información en formato no textuales –imagen, sonido-[77]; y 2) otros términos más específicos, incluidos en el término genérico. Esta compleja estructura de entrada y sub-entradas anidadadas está orientada a la construcción de taxonomías, pero no es adecuada para el diseño de tesauros, que utiliza marcas explícitas para establecer las jerarquías de especialización. Finalmente, el modelo incluye, 3) un conjunto de metadatos que documentan el término.

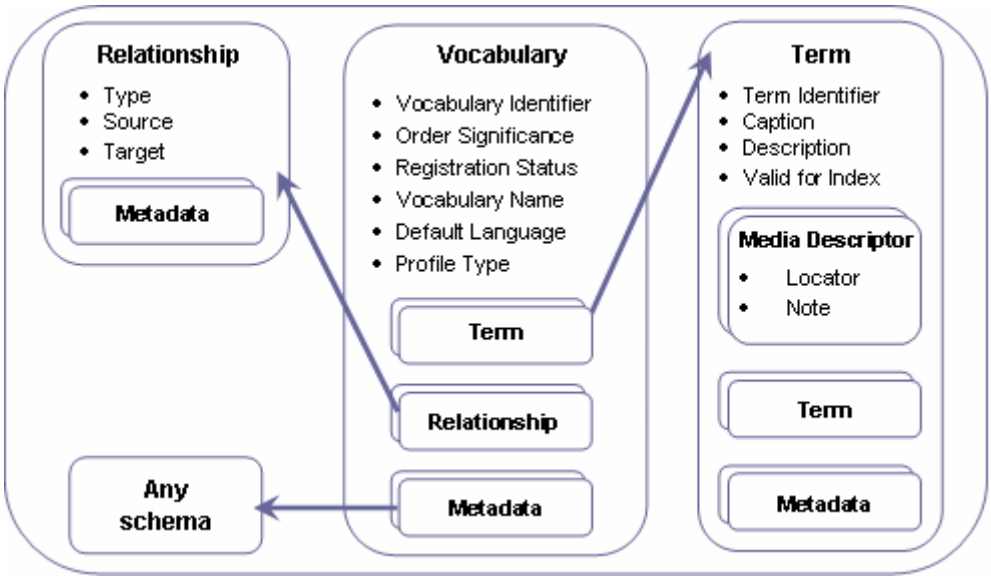


Figura 5.33. Modelo IMS VDEX[78]

[77] La inclusión de otros formatos de definición de términos es una característica única entre los modelos de vocabularios y hace más accesible el contenido del vocabulario a personas con discapacidad visual o auditiva al permitir incluir definiciones en audio, imagen o vídeo.

[78] Fuente wikipedia.org

Cada una de las relaciones entre los términos se define mediante un elemento relación con tres componentes: (i) el término origen, (ii) término destino y (iii) tipo de relación (figura 5.34). Los tipos estándar de relación básicos son: USE, UF, BT, NT y RT[79].

```
<?xml version="1.0" encoding="UTF-8" ?>
- <vdex orderSignificant="false" profileType="thesaurus" xsi:schemaLocation="http://www.imsglobal.org/xsd/imsvdex_v1p
  xmlns="http://www.imsglobal.org/xsd/imsvdex_v1p0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  - <vocabName>
      <langstring language="en">An example thesaurus fragment</langstring>
    </vocabName>
  - <term>
      <termIdentifier>MONOTH_00001</termIdentifier>
    - <caption>
        <langstring language="en">ACOUSTIC BARRIERS</langstring>
      </caption>
    </term>
  - <term>
      <termIdentifier>MONOTH_00002</termIdentifier>
    - <caption>
        <langstring language="en">ACOUSTIC INSULATION</langstring>
      </caption>
    </term>
  + <term>
  - <relationship>
      <sourceTerm>MONOTH_00001</sourceTerm>
      <targetTerm>MONOTH_00002</targetTerm>
      <relationshipType source="http://www.imsglobal.org/vocabularies/iso2788_relations.xml">USE</relationshipType>
    </relationship>
  - <relationship>
      <sourceTerm>MONOTH_00002,</sourceTerm>
      <targetTerm>MONOTH_00001</targetTerm>
      <relationshipType source="http://www.imsglobal.org/vocabularies/iso2788_relations.xml">UF</relationshipType>
    </relationship>
</vdex>
```

Figura 5.34. Ejemplo de tesauro IMS VDEX con tres términos y una relación USE y su inversa UF[80]

Una de las aplicaciones más recientes del IMS VDEX es servir de modelo general para cualquier tesauro que se quiera utilizar en el repositorio federado AGREGA[81] de recursos didácticos para enseñanza primaria y secundaria en España (figura 5.35) (Sarasa, et. al., 2008). Todos los vocabularios controlados del repositorio, listas, taxonomías y tesauros, están construidos con este modelo y en un lenguaje basado en XML, lo que facilita la gestión de sus contenidos, la navegación por la estructura de los vocabularios y la interoperabilidad con otros repositorios de objetos de aprendizaje.

[79] USE y UF son las marcas del estándar ANSI-NISO Z39.19en inglés que se corresponden con las marcas en español USE y USE PARA; BT y NT se corresponde con TG y TE respectivamente; y RT con TR.
[80] Fuente: http://www.imsglobal.org
[81] Disponible en: http://www.proyectoagrega.es/default/Inicio

```
            <caption>
                <langstring>Brainstorming</langstring>
            </caption>
        </term>
        <term>
            <termIdentifier>CAI</termIdentifier>
            <caption>
                <langstring>Computer assisted instruction</langstring>
            </caption>
        </term>
        <term>
            <termIdentifier>cooperativeLearning</termIdentifier>
            <caption>
                <langstring>Cooperative learning</langstring>
            </caption>
        </term>
        <relationship>
            <sourceTerm>brainstorm</sourceTerm>
            <targetTerm>CAI</targetTerm>
        </relationship>
        <relationship>
            <sourceTerm>brainstorm</sourceTerm>
            <targetTerm>cooperativeLearning</targetTerm>
        </relationship>
        <relationship>
            <sourceTerm>CAI</sourceTerm>
            <targetTerm>cooperativeLearning</targetTerm>
        </relationship>
</vdex>
```

Figura 5.35. Aplicación del IMS VDEX al tesauro del repositorio AGREGA[82]

CEN Exchange of Vocabularies (CEN XVD)

El modelo de *intercambio de vocabularios y descripciones*, *eXchange of Vocabularies and Descriptions* (XVD) es un modelo del Comité Europeo para la Estandarización (CEN) cuyo fin es armonizar los vocabularios (CEN CWA 15453, 2005). Consta de dos partes: 1) un modelo conceptual para construir vocabularios multilíngües[83], y 2) un modelo para definir correspondencias entre vocabularios. El modelo conceptual de vocabulario es compatible con IMS VDEX, Zthes[84] y con los estándares de construcción de vocabularios monolingües y multilingües[85].

En el modelo CEN XVD, un tesauro está formado por un conjunto de estructuras denominadas términos equivalentes (termEquivalence) seguidas de un conjunto de

[82] Fuente (Sarasa, et. al, 2008)
[83] El CEN considera vocabularios las listas, taxonomías, tesauros, glosarios y diccionarios.
[84] Zthes es una familia de especificaciones para facilitar la interoperatividad entre aplicaciones que gestionan tesauros (http://zthes.z3950.org/).
[85] ISO 2788 y ANSI/NISO Z39.19 para tesauros monolingües; ISO 5964 para tesauros multilíngües.

estructuras de términos relacionados (figura 5.36). Además incluye cuatro elementos estructurales para documentar globalmente el tesauro: metadatos, general, ciclo de vida, y derechos de autor.

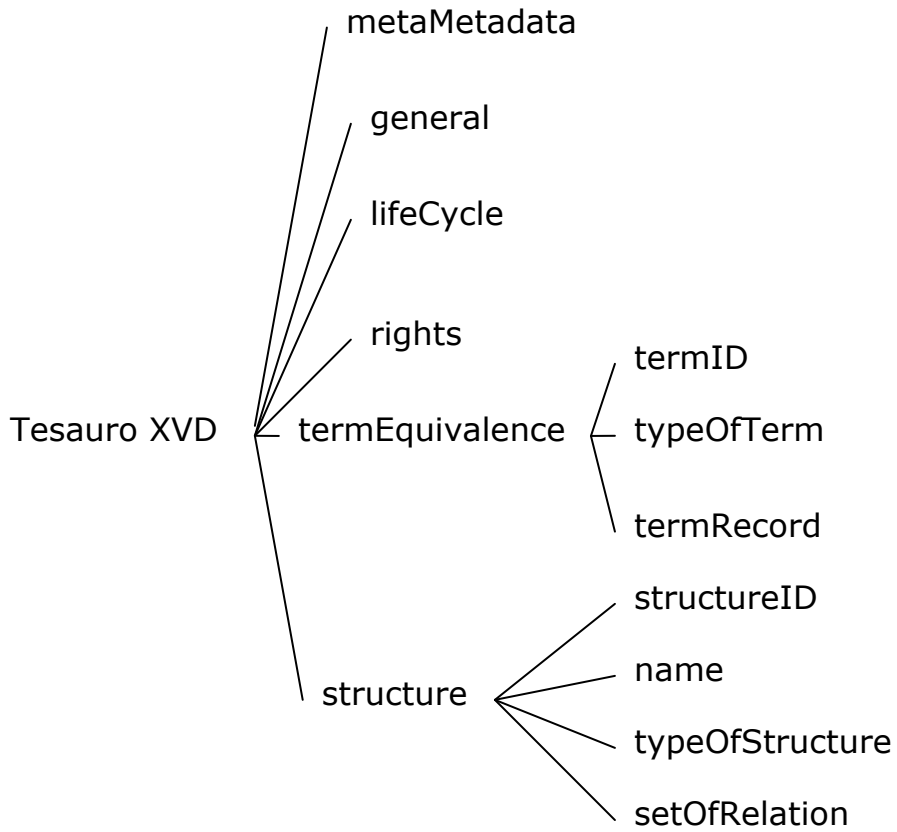


Figura 5.36. Esquema del modelo XVD en XML para tesauros

Los *términos equivalentes* están formadas por: i) un identificador único dentro del vocabulario (termID), ii) el tipo de término (typeOfTerm), que puede ser preferido, no preferido, categoría candidato u obsoleto, y iii) una agrupación (termRecord) con todos los términos equivalentes en una o en distintas lenguas. En la figura 5.37 se muestra, en XML, una estructura de términos equivalentes procedente del tesauro europeo del LRE[86].

El segundo tipo de elementos del modelo XVD son las *estructuras* que definen las relaciones entre los términos y las agrupaciones de términos en categorías. Normalmente, una estructura es una categoría temática o una faceta del tesauro (figura 5.37). Si el tesauro no contiene categorías, estará formado por una única estructura con todos los términos relacionados semánticamente. Cada estructura se identifica con un ID

[86] Learning Resource Exchange Thesaurus, disponible en: http://lre-thesaurus.eun.org/

único en el vocabulario, un nombre y un tipo[87] y contiene el conjunto de relaciones (setOfRelation) del mismo tipo de un término con otros términos. Los tipos de relaciones son los tipos estándares –BT, NT, RT, TT, U, UF-, pero pueden definirse otras relaciones específicas del dominio.

Respecto a las aplicaciones de CEN XVD son escasas (figura 5.38), probablemente debido a que es un modelo bastante reciente y que se añade a otro modelo XML al conjunto de modelos ya existentes en la Red, con la misma capacidad expresiva y con el mismo propósito de servir de herramienta de armonización entre vocabularios.

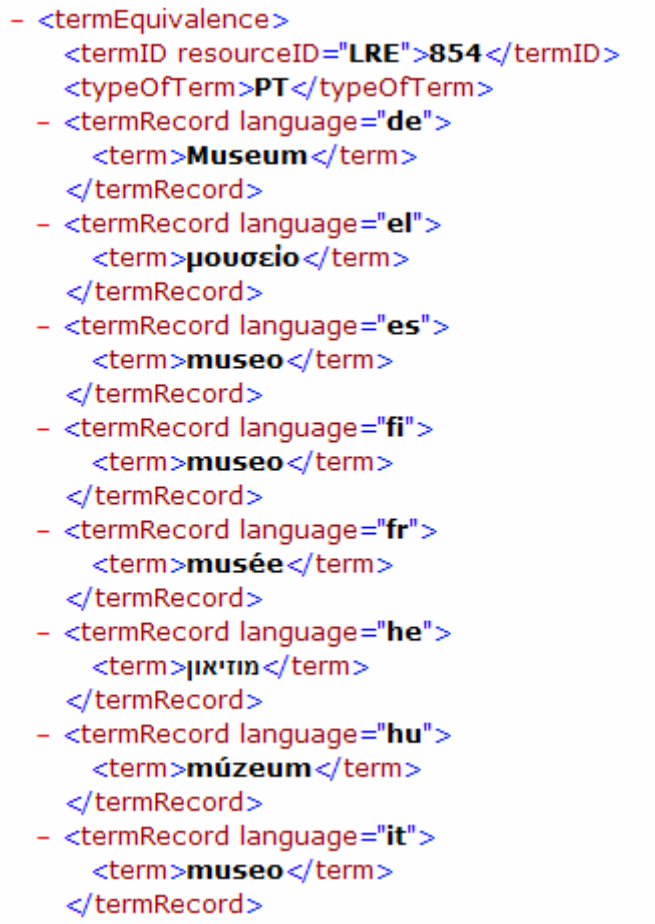


Figura 5.37. La estructura termEquivalence del término preferido (PT) museo (en español) del tesauro LRE[88]

[87] los tipos recomendados son lista, taxonomía, tesauro u ontología, glosario o diccionario.

[88] Fuente http://fire.eun.org/70.xml

```
<structure>
<structureID>130</structureID>
<name>
<string language="da">Miljø</string>
<string language="de">Umwelt</string>
<string language="el">ΠΕΡΙΒΑΛΛΟΝ</string>
<string language="en">Environment</string>
<string language="es">Medio ambiente</string>
<string language="fi">Ympäristö</string>
<string language="fr">Environnement</string>
<string language="he">הביבס</string>
<string language="hu">Környezet</string>
<string language="it">Ambiente naturale</string>
<string language="nl">Milieu</string>
<string language="sv">Miljö</string>
 </name>
<order>alphabetical</order>
<setOfRelation>
<sourceTermID>223</sourceTermID>
<relation>
<typeOfRelation>TT</typeOfRelation>
<order>alphabetical</order>
 </relation>
 </setOfRelation>
<setOfRelation>
<sourceTermID>424</sourceTermID>
<relation>
<typeOfRelation>TT</typeOfRelation>
<order>alphabetical</order>
 </relation>
 </setOfRelation>
<setOfRelation>
<sourceTermID>223</sourceTermID>
<relation>
<typeOfRelation>RT</typeOfRelation>
<order>alphabetical</order>
<targetTermID>224</targetTermID>
<targetTermID>813</targetTermID>
 </relation>
 </setOfRelation>
<setOfRelation>
<sourceTermID>424</sourceTermID>
<relation>
<typeOfRelation>BT</typeOfRelation>
<order>alphabetical</order>
<targetTermID>410</targetTermID>
<targetTermID>1090</targetTermID>
<targetTermID>900</targetTermID>
 </relation>
 </setOfRelation>
</structure>
```

Figura 5.38. Ejemplo de aplicación de VDX para la descripción de la categoría "Medio Ambiente" del tesauro europeo LRE[89].

[89] El término con ID 424, que es raíz de una jerarquía BT cuyos NT son 410, 1090 y 900. Además se

Consideraciones finales sobre los modelos procedentes del e-learning

Los dos modelos de tesauros e-learning, IMS VDEX y CEN VDX, están orientados a la construcción de los vocabularios para los metadatos de los recursos educativos digitalizados, taxonomías y tesauros principalmente. Reproducen los modelos de construcción de los recursos educativos para el *e-learning*, denominados "modelos de contenido de los objetos de aprendizaje"[90]. Los modelos estándares de contenido tratan de construir los contenidos de los objetos de aprendizaje mediante la agregación de contenidos más básicos. El objeto de aprendizaje tiene tres partes: (i) los metadatos, que describen el contenido; (ii) la definición de la estructura del contenido del objeto de aprendizaje, es decir cómo se han agregado los recursos más básicos; y (iii) los contenidos que son los archivos físicos. De la misma forma, los modelos *e-learning* de tesauros consideran que el tesauro se construye de forma estructurada a partir de recursos más básicos, que son, por ejemplo, los términos y las relaciones en el modelo IMS VDEX, o los conjuntos de términos equivalentes, llamados 'termEquivalence', y las estructuras que combinan los conjuntos de términos equivalentes, llamadas 'structure', en el modelo CEN VXD. A su vez, cada una de estos recursos puede descomponerse en otros recursos más simples, como por ejemplo, los términos del modelo IMS VDEX que se descomponen en 'Media Descriptor', 'Term' y 'Metadata' (figura 5.33). En definitiva, los tesauros se organizan igual que los recursos educativos con el objetivo de unificar el tratamiento automático, y que las mismas herramientas que gestionan los recursos educativos sirvan para gestionar el contenido de los tesauros. Desde este punto de vista, puede decirse que los tesauros son también recursos educativos u objetos de aprendizaje, que se utilizan para construir otros objetos cuando se integran en los metadatos. La diferencia con los objetos de aprendizaje está en que la estructura de los tesauros ya está predefinida -y de forma muy diferente en los dos estándares IMS VDEX y CEN VXD. La cuestión que queda abierta es porqué no dejar que los tesauros tengan su propia estructura, siempre que dicha estructura se defina siguiendo el modelo estándar de contenido para garantizar la interoperabilidad.

---

muestra la relación de asociatividad entre el término con ID 223 y los términos 224 y 813. Los lemas asociados a estos ID se definen en la estructura termEquivalence.

[90] Los estándares SCORM Content Aggregation Model disponibles en: http://xml.coverpages.org/SCORM-12-CAM.pdf, e IMS Content Packaging, disponible en: http://www.imsglobal.org/content/packaging/

5.4. Resumen y conclusiones del capítulo

Los modelos de datos informáticos utilizados conjuntamente con el modelo de los estándares de construcción de tesauros (capítulo 4) proporcionan, frente a los modelos tradicionales sistemático y alfabético, dos ventajas importantes: 1) una metodología de construcción y gestión basada en la independencia entre el contenido, la estructura y la presentación del tesauro que a) facilita el mantenimiento de la consistencia de los datos y b) amplía las posibilidades de acceso al contenido; 2) el tratamiento automático del contenido y, en consecuencia, la mejora de la eficacia y eficiencia del tesauro como sistema de representación y recuperación de la información o de los recursos digitalizados de un dominio.

Los modelos de datos conceptuales sirven para describir, a nivel lógico, la organización del contenido, mientras que los modelos de implementación de datos organizan el contenido en estructuras de datos que son procesables por las aplicaciones informáticas. Normalmente se utilizan ambos tipos de modelos para diseñar los tesauros en dos fases: en la primera, que es la del diseño conceptual, se obtiene el esquema lógico del tesauro, y en la segunda, de implementación, se traduce a un esquema de implementación de datos. Los modelos para el diseño conceptual revisados en este capítulo son los modelos basados en grafos (i) redes semánticas e hipertexto, (ii) el modelo Entidad-Relación y Entidad-Relación Extendido, y (iii) el modelo orientado a objetos y UML. Estos modelos conceptuales tienen dos limitaciones para representar de forma natural ciertas estructuras del contenido estándar de los tesauros: 1) no hay un mecanismo para definir las categorías como conjuntos de términos, en el sentido matemático y con las operaciones y propiedades de los conjuntos, y a la vez, para definir las relaciones entre categorías y términos como elementos de un grafo, a los que también se les puedan aplicar las operaciones y propiedades de los grafos; 2) no es fácil definir de forma general las relaciones asociativas simétricas, como la relación TR, entre varios términos, llamadas también hiperarcos. Aunque existen modos de resolver estas limitaciones, los esquemas resultantes no son generales, dependen del ámbito de cada tesauro, y son poco flexibles a los cambios de estructura. Esto significa que en ciertos casos se tendría que rehacer el esquema lógico y, en consecuencia, rehacer el tesauro, con los esquemas de implementación de datos y los datos ya almacenados.

Los modelos de implementación de datos utilizados para construir tesauros son: (i) el modelo relacional y (ii) los modelos basados en XML. El modelo relacional construye el tesauro en una base de datos relacional, formada por conjuntos de relaciones o tablas que están relacionadas entre sí por los valores de algunas de sus columnas. Es el modelo de implementación más utilizado porque dispone de 1) un sólido fundamento teórico para diseñar bases de datos muy optimizadas, con una redundancia de datos mínima y con un buen sistema para controlar automáticamente la consistencia de los datos; 2) lenguajes declarativos, relativamente fáciles de utilizar, para la creación y gestión, consulta y actualización, de los datos; y 3) un amplio abanico de aplicaciones para su gestión. El inconveniente de este modelo es la dificultad para gestionar estructuras de datos jerárquicas, como las jerarquías TG-TE de términos de los tesauros, aunque es una cuestión que en las últimas versiones de los lenguajes de gestión de datos ya está resuelto.

Los modelos de datos basados en XML construyen los tesauros en uno o varios archivos de texto con el contenido del tesauro etiquetado. El lenguaje de etiquetas sirve para definir y distinguir los componentes del tesauro, sus propiedades y sus restricciones. No existe un lenguaje de etiquetas único, sino varias propuestas de lenguajes particulares y estándares. Cada uno de estos lenguajes está formalmente definido con el metalenguaje XML en gramáticas independientes del contexto. Actualmente, las propuestas estándar para tesauros XML son de dos tipos: (i) basadas en el modelo RDF, y (ii) basadas en el modelo de contenido de objetos *e-learning*. En el primer caso, RDF, el tesauro se describe como un conjunto de objetos con propiedades y valores. La extensión de RDF, RDFS, permite, además, definir clases de objetos y restricciones sobre el tipo de objetos y las clases. El modelo OWL extiende RDFS añadiendo la posibilidad de expresar relaciones entre las clases, restricciones de cardinalidad de las relaciones, operaciones booleanas y operaciones de conjuntos para realizar inferencias. El modelo SKOS-Core es, por su parte, una simplificación de OWL, orientada a la representación de estructuras conceptuales como las ontologías y los tesauros ontológicos. El objetivo de los modelos basados en RDF es la representación unificada de estructuras conceptuales, tesauros y ontologías principalmente, para que sirvan de modelos semánticos en el procesamiento inteligente de los recursos y de la información contenida en la Web.

Los modelos de datos procedentes del *e-learning*, IMS VDEX y CEN XVD, están orientados a la construcción de los vocabularios que se integren en los metadatos de los recursos educativos. Estos modelos se basan en los modelos de construcción de los

objetos de aprendizaje *e-learning*: (i) el contenido es la agregación de componentes básicos o de otros objetos; (ii) el objeto tiene una definición de la estructura de agregación; y (iii) el objeto se describe con metadatos. La diferencia entre los dos modelos, IMS VDEX y CEN XVD, es la estructura o forma de agregación de los componentes del tesauro. La diferencia con los objetos de aprendizaje, por su parte, es que, en éstos últimos, el autor del objeto puede decidir cómo quiere agregar los componentes y en el caso de los modelos de vocabularios, IMS VDEX y CEN XVD, la forma de agregación está ya definida y es fija. El propósito de los modelos *e-learning* es la representación unificada de los vocabularios y de objetos de aprendizaje para que ambos puedan ser gestionados en un mismo entorno tecnológico, por ejemplo, en un repositorio de objetos de aprendizaje.

Paradójicamente, cualquiera de los tipos de modelos XML presentados, basados en RDF o en *e-learning*, tiene como objetivo servir de marco común para facilitar la interoperabilidad y la reutilización del contenido de los tesauros. Los modelos RDF están basados en una concepción ontológica de los tesauros, mientras que los modelos basados en *e-learning* siguen una concepción más terminológica. Todos ellos tienen suficiente capacidad de representación, pero la eficiencia está siendo cuestionada. Además, aunque son modelos extensibles, son rígidos, porque imponen una concepción y estructura del tesauro determinada. Los modelos XML se consideran, actualmente, formatos de intercambio de datos, pero no son modelos adecuados para un diseño flexible y un procesamiento eficiente de la información.

En definitiva, para la construcción de tesauros de explotación con aplicación al *e-learning*, se debería aprovechar la capacidad expresiva de los modelos de vocabularios XML del *e-learning*, y la uniformidad con los modelos de contenido para garantizar un tratamiento informático común en plataformas *e-learning*, pero al mismo tiempo, no es deseable la rigidez innecesaria, de una estructura de agregación fija. Se trata de ir hacia un planteamiento más abierto *buscando un modelo conceptual suficientemente expresivo, uniforme y flexible* como para diseñar esquemas de datos que se *adapten a cualquier estructura y contenido de los tesauros*, pero que, al mismo tiempo, dicha estructura se pueda definir utilizando un modelo de implementación de datos estándar, relacional o basado en XML, para garantizar, cuando sea necesario *la eficiencia*, aplicando el modelo relacional, *o la interoperabilidad*, aplicando un modelo estándar XML.

Capítulo 6

El modelo higraph léxico para la construcción de los tesauros

If I can't picture it, I can't understand it (A. Einstein)

Este capítulo propone un nuevo modelo de datos, el *higraph léxico*, para la representación sistemática y visual del contenido de los tesauros. Este modelo tiene la capacidad de describir las estructuras de organización del conocimiento léxico con independencia del tipo de tesauro, del tipo de ámbito, del tipo de estructura sistemática, del tipo de contenido y del tipo de actualización que pueda experimentar. Es un modelo general y flexible que puede reajustarse y recoger los cambios de la evolución no determinista del léxico de las lenguas, que no depende de unas condiciones iniciales. Tiene un carácter matemático y lingüístico, puesto que se obtiene aplicando el formalismo matemático topo-visual de los *higraphs,* introducidos por D. Harel en el año 1988 a la representación del dominio léxico como un sistema lingüístico de corte estructural. El resultado es, en realidad, un metamodelo[1] que sirve para construir los modelos específicos de tesauros en sus aspectos estructural y operacional. Aunque preferimos la denominación de modelo, en vez de metamodelo, de acuerdo con una de las tres definiciones propuestas por Pidcock (2003): un metamodelo es "…un modelo de un dominio de interés". El carácter de metamodelo aporta a los procesos de construcción y explotación de los tesauros: (i) una arquitectura clara, uniforme y general, (ii) una implementación eficiente, (iii) escalabilidad y (iv) mayor flexibilidad (Bezivin, 2005).

Este capítulo presenta el modelo de higraph léxico siguiendo el esquema siguiente: en la primera sección, se introduce el modelo matemático-visual de higraph; en la segunda sección, se resume el modelo lingüístico de significado de los términos de la lengua; en la tercera sección, se aplica el modelo de higraph a los tesauros, definiendo una interpretación para los elementos del higraph acorde con el significado de los términos de las lenguas; en la cuarta sección, se propone una implementación del modelo de higraph léxico utilizando bases de datos relacionales; finalmente, en la quinta sección, se hace un resumen del modelo propuesto.

[1] Un metamodelo es un modelo explícito de constructores y reglas para crear modelos específicos de los dominios de interés (Pidcock, 2003).

6.1. El modelo matemático y visual de los higraphs

El modelo de *higraphs* (H) fue propuesto por David Harel (Harel, 1988:1) para representar "información de naturaleza compleja y complicada […], no cuantitativa, organizada con interrelaciones y agregaciones […]". Es un formalismo visual, con una sintaxis y semántica definida de forma precisa que es fácilmente interpretable por las personas y las máquinas[2]. Los higraphs se pueden expresar en otros formalismos de representación del conocimiento con el fin procesarlos y manipularlos automáticamente. Zadrony (1991) presenta una correspondencia con el lenguaje Datalog, que es un lenguaje de bases de datos basado en la lógica de predicados para representar los datos y las reglas de inferencia. Paige (1995) traduce los H a la Lógica de Predicados para su aplicación en la especificación y verificación formal de algoritmos. Fogarty (2006) presenta una correspondencia entre los H y los Grafos Dirigidos Acíclicos, para aplicarlos al modelado de sistemas de ingeniería. Siguiendo esta línea de trabajos, este capítulo presenta, en la sección cuarta, una correspondencia entre los H y higraphs léxicos con el modelo de bases de datos relacional para aplicarlos a la construcción y gestión de tesauros monolingües.

6.1.1. Sintaxis

El modelo de higraphs combina las teorías de grafos y diagramas de Venn para representar estructuras compuestas por conjuntos, jerarquías de conjuntos y subconjuntos, conjuntos ortogonales y relaciones entre conjuntos (de aridad en general n)[3]. Formalmente un higraph se define con cuatro elementos H = (G, σ, π, E). El primer componente, G, es un conjunto finito de glóbulos[4]. Un glóbulo es un conjunto pero con elementos que son otros glóbulos o nada (∅). Se representan con diagramas de Venn. Las operaciones de conjuntos de unión, intersección, diferencia y producto cartesiano, no ordenado, pueden aplicarse a los glóbulos. Los glóbulos son una extensión de los conjuntos, para que admitan un significado más amplio: cualquier estructura que englobe a otras. Los glóbulos *atómicos* no incluyen ningún otro glóbulo, son los elementos indivisibles del modelo. El resto de los glóbulos pueden ser divididos en subglóbulos mediante la función σ y la función π. Esto significa que los glóbulos no

[2] Por ejemplo, *Structure 101* es una herramienta comercial utilizada para definir y gestionar higraphs aplicados al desarrollo de software: http://www.headwaysoftware.com/products/structure101/index.php
[3] Los grafos de un higraph son *hipergrafos.*
[4] David Harel los denomina *Blobs*.

elementales pueden contener otros glóbulos, formando jerarquías de inclusión (σ) o bien formando particiones ortogonales si son disjuntos[5] (π).

La función σ define los subglóbulos, no necesariamente disjuntos, pertenecientes a un glóbulo de G:

$$\sigma: G \rightarrow 2^G$$

Donde 2^G el conjunto de las partes de G, es decir, todos los posibles subglóbulos que se pueden formar con los componentes de G

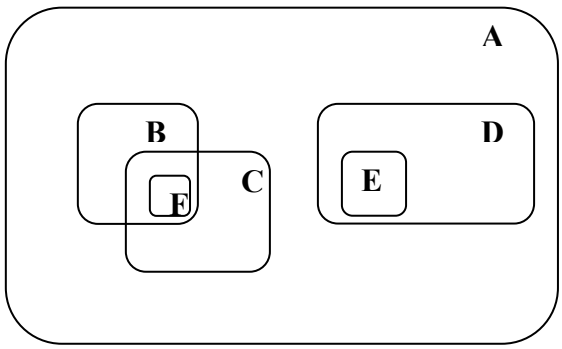


Figura 6.1. Ejemplo de una jerarquía de glóbulos

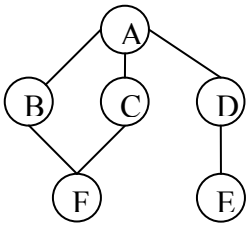
La figura 6.1, muestra un posible ejemplo de una jerarquía de glóbulos que puede definirse con una secuencia de inclusiones:

$$A \supset B \supset F \supset \varnothing \text{ y } A \supset C \supset F \supset \varnothing \text{ y } A \supset D \supset E \supset \varnothing$$

o bien, con la función sigma:

$$\sigma(A)=\{B,C,D\} \text{ y } \sigma(B)= F \text{ y } \sigma(C)= F \text{ y } \sigma(D)= \{E\} \text{ y } \sigma(E)= \varnothing \text{ y } \sigma(F)= \varnothing$$

O bien, gráficamente:



El conjunto de *glóbulos atómicos* de la figura 6.1 es Atom={F,E}. La intersección de dos glóbulos es siempre otro glóbulo o conjunto vacío (∅): $B \cap C = F$ y F, B, C $\in$ A

La *unión* es la operación inversa de la descomposición. Construye un glóbulo a partir de los subglóbulos contenidos en un mismo nivel de inclusión. Por ejemplo, el glóbulo A se forma con la unión de los subglóbulos del nivel 1 B, C y D:

[5] No es correcto utilizar el concepto de conjuntos disjuntos porque supone que el glóbulo contiene elementos, pero hace más sencilla la presentación de los nuevos conceptos.

$$A = B \cup C \cup D$$

Si se representa la unión de los subglóbulos de cada nivel de la forma $\sigma^{nivel}(A)$, entonces la composición de un glóbulo, incluyendo todos los niveles de profundidad de la jerarquía de inclusión, es *el cierre de la función sigma*:

$$\sigma^{+}(A) = \cup \{\text{para i desde 1 hasta } \infty\}.\ \sigma^{i}(A) \text{ y tal que } A \notin \sigma^{+}(A)$$

Así, el cierre del glóbulo A de la figura 6.1, se calcula de la forma siguiente:

$$\sigma^{+}(A) = \cup \{\text{para i desde 1 hasta 2}\}.\ \sigma^{i}(A) = \sigma^{1}(A) \cup \sigma^{2}(A) = \{B,C,D\} \cup \{E, F\} =$$
$$= \{B, C, D, E, F\}$$

Otra manera de separar un glóbulo en sus componentes, que son también glóbulos, es mediante la *partición ortogonal* de un glóbulo en partes disjuntas. La *función π*, aplicada a un glóbulo, calcula los subglóbulos *disjuntos* (definidos por σ(X)) de X o, dicho de otra manera, divide el glóbulo X en clases de equivalencia o en particiones ortogonales que no comparten ningún glóbulo en común ya que la intersección de las partes ortogonales es vacía. En general:

$$\pi: G \rightarrow 2^{GxG}$$

Donde G es el conjunto de los glóbulos y 2^{GxG} es el conjunto de todos los posibles pares de subglóbulos que se pueden formar con los componentes de G.

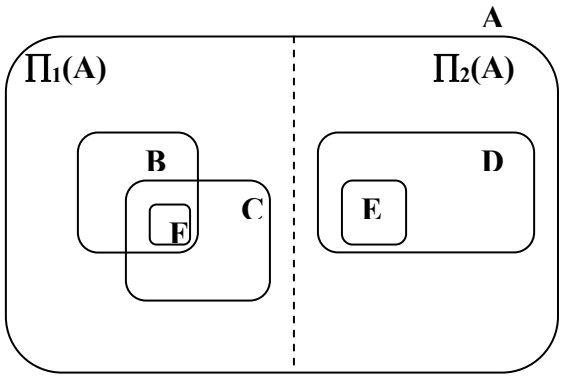


Figura 6.2. Una posible descomposición ortogonal del glóbulo A

En la figura 6.2 se han separado los componentes de A en dos glóbulos disjuntos, $\pi_1(A)$ y $\pi_2(A)$, el primero contiene a B y C, y el segundo a D:

$$\pi(A) = (\pi_1(A), \pi_2(A)) \text{ y } \pi_1(A)=\{B,C\} \text{ y } \pi_2(A)=\{D\}$$

La intersección entre el contenido de ambas partes es vacía: $\sigma^{+}(\pi_1(A)) \cap \sigma^{+}(\pi_2(A)) = \varnothing$

La operación inversa a la descomposición ortogonal es el *producto cartesiano*, que compone las partes en un glóbulo en otro glóbulo superior. Este producto cartesiano

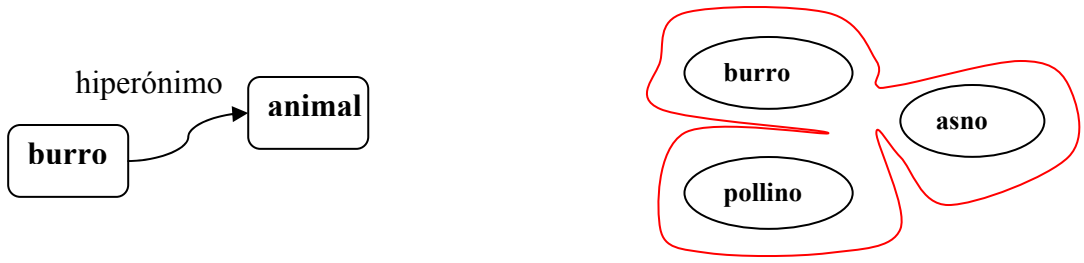
presenta una diferencia con respecto al producto cartesiano de la teoría de conjuntos: no es ordenado, los glóbulos operandos no tienen un orden definido. Formalmente, el producto cartesiano es el conjunto tuplas[6] o lista de glóbulos pertenecientes a cada parte ortogonal:

$$A = \pi_1(A) \otimes \pi_2(A) = B \cup C \otimes D = \{(m,n) \mid m \in B \cup C \text{ y } n \in D\}$$

Finalmente, el último componente de un higraph es el conjunto de arcos, E, que representan las conexiones o asociaciones entre glóbulos. Los arcos, en teoría de grafos, se definen con el conjunto de los nodos que conectan y se representan por tuplas.

Los arcos que conectan dos elementos son arcos binarios, cuando conectan n elementos son arcos n-arios. La sinonimia total es un ejemplo de este tipo de relaciones en las que todos los elementos están conectados con todos y las conexiones son simétricas. Se denominan también, hiperarcos, y se representan por una tupla de tamaño n (figura 6.3 derecha).

Los arcos pueden tener asociada una etiqueta, que representa información sobre la relación, como el tipo de relación. En ese caso se añade la etiqueta a la tupla con los nodos. Si la relación es asimétrica el arco tiene un sentido, se dice dirigido, y se representa mediante una tupla ordenada, por ejemplo, (nodo_origen, nodo_destino) (figura 6.3 izquierda).



Arco: (hiperónimo, burro, animal) Hiperarco: (sinonimia, burro, asno, pollino)

Figura 6.3. A la izquierda, ejemplo de arco binario y etiquetado con el tipo de asociación, hiperonimia, entre los glóbulos "burro" y "animal". A la derecha, ejemplo de hiperarco para representar la sinonimia.

En un higraph los arcos e hiperacos conectan los glóbulos. En la figura 6.4 se han incorporado un conjunto de arcos binarios, E, al higraph de la figura 6.2, siendo E:

$$E = \{(D,A), (B,C), (F,E)\}$$

[6] En este caso, las tuplas no están ordenadas, es decir, que son, simplemente, una secuencia de glóbulos, cada uno de ellos perteneciente a una de las partes ortogonales que se están "multiplicando".

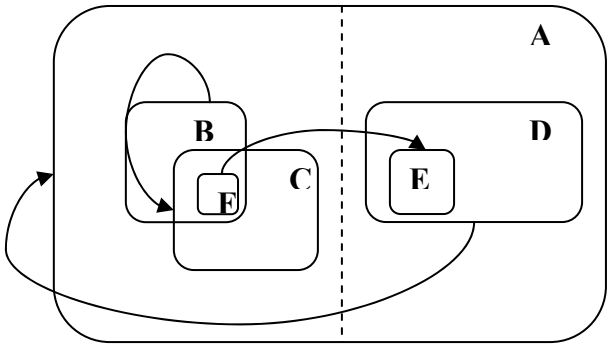


Figura 6.4. La estructura de glóbulos con los arcos forma un higraph completo

6.1.2. Semántica

La semántica de un higraph se define composicionalmente a partir del significado de los glóbulos atómicos. El significado de los glóbulos atómicos se define, a su vez, mediante una función de interpretación, μ, que asigna a cada glóbulo atómico un valor de significado de un conjunto no estructurado[7], D, de significados disjuntos. Si se llama Atom al conjunto de glóbulos atómicos de un higraph; la función de interpretación, μ, se define formalmente como:

$$\mu: \text{Atom} \rightarrow 2^D$$

y cumple la restricción de que dos glóbulos atómicos distintos no pueden tener ningún significado en común:

$$\text{Si } X \neq Y \text{ entonces } \mu(X) \cap \mu(Y) = \varnothing$$

El significado de los glóbulos complejos depende del significado de los glóbulos atómicos que los componen y de cómo se combinan estos significados de forma incremental, utilizando las operaciones de composición sintáctica unión y producto cartesiano antes definidas:

$$\mu(Z) = \otimes_{(\forall i \in \{1..k\})} \left(\cup_{(\forall P \in \pi i(Z))} \mu(P) \right)$$

Así, el significado de $Z \in G$ se calcula concatenando, con el operador producto cartesiano, los significados de cada una de sus componentes ortogonales, y estos componentes, a su vez, se interpretan uniendo cada una de las interpretaciones de los glóbulos que lo constituyen.

Si no existen particiones en el glóbulo Z, la definición del significado se simplifica a la forma:

$$\mu(Z) = \cup_{(\forall P \in \sigma(Z))} \mu(P)$$

[7] Se considera que D no es estructurado para evitar tener un elemento x, y también {x}.

Finalmente, se define el significado de los arcos. Un arco induce una relación semántica en los significados de los glóbulos conectados. Cuando hay un arco entre dos glóbulos, X e Y, este arco representa una relación semántica entre los significados de X e Y:

$$(\mu(X), \mu(Y)) \in E_M \text{ si y sólo si } (X,Y) \in E$$

Para aplicar el modelo semántico a un dominio de información, se define, en primer lugar, el dominio de valores del significado, D, y la función μ, que asocia los elementos atómicos del higraph con el conjunto de valores de D. En segundo lugar, se construye incrementalmente el significado de los glóbulos no atómicos hasta obtener la interpretación de todo el higraph. En el dominio de las lenguas naturales y de los tesauros, el modelo higraph aporta una formalización y visualización de la naturaleza preferente de las relaciones semánticas clásicas de inclusión y partición del lenguaje natural y, también, del resto de relaciones. La aplicación del modelo es directa si se define el conjunto D como el conjunto de valores del significado en un sistema de signos. Esta es la hipótesis de la que partimos para demostrar que los tesauros pueden estructurarse matemática y visualmente con el modelo de higraphs y pueden interpretarse con el modelo lingüístico de sistema autónomo de signos.

6.2. El tesauro como un sistema autónomo de signos

Ferdinand Saussure consideraba que la lengua es un sistema estructurado de signos en el que el *valor* del significado de cada elemento dependía de su posición diferencial respecto de los demás y no de su valor intrínseco (Saussure, 1916). El modelo de higraphs, además, representa formal y visualmente sistemas con estructuras complejas, altamente relacionadas. Ambos hechos nos sirven de punto de partida para formular una nueva aplicación de los higraphs que denominamos *higraph léxico*, que es una de las aportaciones de este trabajo de tesis.

La hipótesis formulada al comienzo de esta memoria (capítulo 1, sección 1.4.2), se puede concretar utilizando el modelo de higraph léxico de la forma siguiente: si 1) se considera que los tesauros monolingües o multilingües son representaciones parciales de una lengua o de varias, y 2) que la lengua es un sistema estructurado de signos relacionados, y 3) que los higraphs son modelos para la representación de sistemas estructurados complejos, entonces los tesauros son subsistemas de signos de la lengua (o lenguas) que pueden representarse matemáticamente y visualmente con el formalismo de los higraphs.

En la próxima sección se demuestra esta hipótesis aplicando los higraphs al contenido estándar de los tesauros (capítulo 4). El resultado de esta aplicación es la descripción directa de la estructura de los tesauros en forma de higraph. La interpretación semántica, sin embargo, se realiza desde una perspectiva lingüística que considera que el significado de cada elemento del tesauro es extraindividual, de existencia sólo social, dependiente de las correspondencias y oposiciones entre sus elementos[8].

Para explicar la demostración se va a utilizar el tesauro de la figura 6.5a. Esta figura muestra la presentación sistemática tradicional de un hipotético tesauro de animales diseñado para que sirva de muestra, ya que contiene todos los tipos de elementos definidos en los estándares de tesauros (categorías, términos y relaciones TG/TE, USE/UP, TR) y varios tipos posibles de estructuras de un higraph. Como se puede comprobar, cualquier término se define a partir de sus relaciones con los otros términos. Así, por ejemplo, el término *reptil* es un término específico –marcado con TE- del término *animal*, que, a su vez, es un término genérico del término *ofidio*. Asimismo, *reptil* está incluido en la categoría EUROPEO y esta categoría está incluida en la faceta ANIMAL SALVAJE[9].

[8] Amado Alonso: Prólogo a la edición española del Curso de Lingüística General de F. Saussure.

[9] Notación: se utilizan minúsculas para indicar términos y las mayúsculas para indicar categorías o facetas, que son conjuntos de términos.

ANIMAL SALVAJE
animal
TE mamífero
TE bóvido
TE muflón
TE lobo
TE reptil
TE ofidio
irracional
USE animal
bicho
USE animal
EUROPEO
tortuga
mamífero
TE bóvido
TE muflón
TE lobo
reptil
TE ofidio
ofidio
lobo
bóvido
TE muflón
muflón
AMERICANO
tortuga (EUROPEO)
ANIMAL DOMÉSTICO
guarro
TR puerco
TR cerdo
puerco
TR guarro
TR cerdo
chancho
USE cerdo
UNIVERSAL
cerdo
TE cerdo ibérico
TR guarro
TR puerco
ESPAÑA
cerdo ibérico

Figura 6.5 a. Presentación sistemática del tesauro de muestra

6.3. El modelo de higraph léxico para tesauros

La representación gráfica del tesauro de la figura 6.5.a se muestra en la figura 6.5.b y es el resultado de establecer una correspondencia, a nivel sintáctico, entre los elementos constituyentes de un tesauro –definidos en los modelos estándares de construcción de

tesauros[10]- y los elementos de un higraph. La correspondencia a nivel semántico se define en base al concepto lingüístico de *valor del significado* de los términos aplicando el modelo composicional de los higraphs.

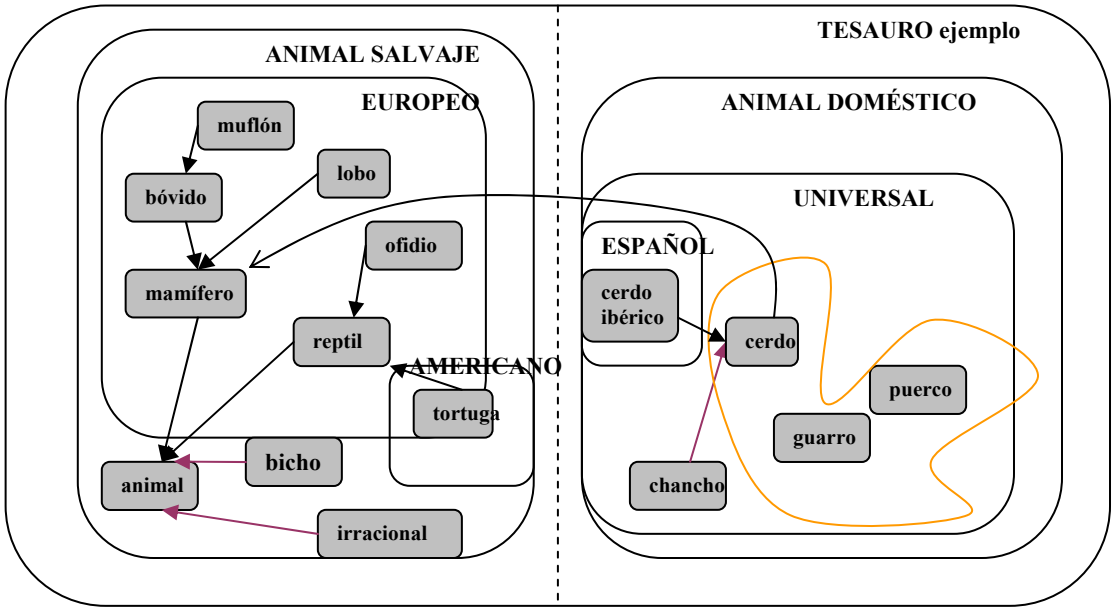


Figura 6.5.b. Higraph léxico[11] del tesauro 6.5.a.

6.3.1. Sintaxis

Un higraph léxico (HL) es una cuaterna:

$$HL = (S, \sigma, \pi, R)$$

donde S es el conjunto de todos los signos que en el tesauro son los términos y categorías (temáticas y facetas[12]); sigma (σ) es la función que calcula la jerarquía de categorías y de pertenencia de términos a categorías; la función π calcula las facetas y R es el conjunto de relaciones semánticas del tesauro, que consideraremos que son las relaciones estándar TG/TE, USE/UP, TR, sin que esto sea óbice para extenderlas cuando sea necesario con otro tipo de relaciones específicas del dominio de conocimiento del tesauro.

Se define la correspondencia entre (S, σ, π, R) y los componentes de un higraph (G, σ, π, E), de la forma siguiente:

[10] En concreto se utiliza la versión española (UNE 50106, 1990) del estándar (ANSI/NISO Z39.19, 2005), pero estos elementos son comunes a todos los estándares de construcción de tesauros monolingües.
[11] Los arcos en **negro** son relaciones TG/TE –apuntando al TG-, los arcos en **morado** son U/UP –apuntando al término preferido-, y la línea **naranja** denota un hiperarco TR, con tres términos cuasi-sinónimos. Todas las relaciones están referidas al significado 'animal'.
[12] Las facetas son categorías, que representan cada una de las dimensiones de un dominio (capítulo 4).

1. los términos del tesauro son los glóbulos atómicos del HL, porque no contienen ningún otro elemento del tesauro[13];
2. las categorías son glóbulos no atómicos del HL porque contienen términos u otras categorías. Las facetas también son categorías pero no contienen ningún término o categoría común entre ellas;
3. los signos del tesauro, S, son los glóbulos (atómicos y complejos) del higraph, y en el HL son todos los términos y las categorías del tesauro;
4. las categorías y facetas contienen términos y también pueden contener a otras categorías formando la jerarquía σ del HL;
5. las categorías pueden dividirse en categorías disjuntas, que en la terminología de tesauros se denominan facetas. Las facetas son las particiones ortogonales de un HL definidas por la función π;
6. Existe una única categoría raíz que contiene a todas las demás y que es el tesauro.
7. las operaciones de conjuntos de los higraphs –unión, intersección, diferencia y producto cartesiano no ordenado- están definidas igualmente en un HL, así como la relación de inclusión (o "miembro de") entre categorías y otras categorías y términos;
8. las relaciones del tesauro se representa con los arcos del HL, pero etiquetados, porque en los tesauros, las relaciones son de un tipo determinado, -los tipos estándares son TG/TE, USE/USEPARA y TR[14-]. Esto significa que los arcos además de los nodos, tienen un componente más que es la etiqueta que indica el tipo de relación que asocia a los nodos del arco.

El tesauro ejemplo 6.5.a, representado visualmente en la figura 6.5.b, es un HL que, aplicando las correspondencias anteriores, se define matemáticamente con la cuaterna:

Tesauro ejemplo = (S, σ, π, R)

S = Categorías $\cup$ T

Categorías = {Tesauro ejemplo, ANIMAL SALVAJE, ANIMAL DOMÉSTICO, EUROPEO, AMERICANO, UNIVERSAL, ESPAÑOL};

T = {t | t es un término del tesauro} = {animal, mamífero,…, cerdo ibérico}[15]

[13] Para simplifcar la presentación no se van a considerar las notas de ámbito. En caso de existir, se consideran igual que si fueran relaciones específicas del dominio. El de tipo de relación sería NA entre el término y otro glóbulo que contiene el texto de la nota.

[14] TG: Término General; TE: Término Específico, inverso de TG; USE: Usar término preferido; USE PARA: inversa de USE; TR: Término Relacionado.

[15] Los términos se escriben con minúsculas para distinguirlos de las categorías.

σ(ANIMAL SALVAJE) = {EUROPEO, AMERICANO, animal, bicho, irracional}

σ(ANIMAL DOMÉSTICO) = {UNIVERSAL}

σ(animal) = $\varnothing$; …; σ(cerdo ibérico) = $\varnothing$

π1(TESAURO ejemplo) = ANIMAL SALVAJE;

π2(TESAURO ejemplo) = ANIMAL DOMÉSTICO;

R = {(TG, mamífero, animal)[16], (TG, reptil, animal), (TG, ofidio, reptil), (TG, mamífero, lobo), (TG, mamífero, bóvido), (TG, bóvido, muflón), (TG, tortuga, reptil), (USE, irracional, animal)[17], (USE, bicho, animal), (TG, cerdo ibérico, cerdo), (TG, cerdo, mamífero), (USE, chancho, guarro), (TR, cerdo, puerco, guarro)}[18]

Las operaciones de conjuntos y grafos en el HL *Tesauro ejemplo* permiten representar gráficamente y manipular automáticamente la información contenida en el tesauro. Por ejemplo, las operaciones de conjuntos:

Unión: Obtener la información de la categoría "Animal salvaje". Es la unión de todos sus componentes, términos y subcategorías:

ANIMAL SALVAJE = EUROPEO $\cup$ AMERICANO $\cup$ animal $\cup$ bicho $\cup$ irracional = {muflón, bóvido, mamífero, animal, lobo, ofidio, reptil, tortuga, bicho, irracional}

Intersección: Calcular qué animales son "Europeos" y "Americanos":

EUROPEO $\cap$ AMERICANO = tortuga

Diferencia: Calcular los animales domésticos que no son españoles:

UNIVERSAL – ESPAÑOL = {cerdo, chancho, guarro, puerco}

Producto cartesiano no ordenado: Calcular los posibles pares de "animales salvajes" con "animales domésticos":

ANIMAL SALVAJE $\otimes$ ANIMAL DOMÉSTICO = {(término1, término2) | (término1 $\in$ ANIMAL SALVAJE ó termino1 $\in$ ANIMAL DOMÉSTICO) y (término2 $\in$ ANIMAL SALVAJE ó término2 $\in$ ANIMAL DOMÉSTICO}

Por ejemplo: (muflón, cerdo ibérico) $\in$ ANIMAL SALVAJE $\otimes$ ANIMAL DOMÉSTICO

Las operaciones sobre grafos permiten manipular automáticamente la red de relaciones semánticas terminológicas para insertar nuevos términos, borrar, buscar, actualizar y cualquier consulta que implique recorrer la red. Por ejemplo, se podría consultar si el

[16] Esta relación se lee "el Término Genérico de mamífero es animal".
[17] Esta relación USE se lee "USE animal para irracional".
[18] No se representan las relaciones inversas TE (de TG), USE PARA (de USE), porque pueden obtenerse simplemente cambiando el orden de lectura de los elementos de la relación. Por ejemplo, (TG, mamífero, animal) se podría leer como: "el TE de animal es mamífero" o bien "el TG de mamífero es animal".

"cerdo ibérico" es un "mamífero". Como existe un camino entre ambos términos de arcos TG (ó TE), la respuesta es afirmativa.

6.3.2 Semántica

Sabiendo que se pueden definir varios modelos M= (D, μ)), nosotros proponemos una semántica del HL teniendo en cuenta (sección 6.2) que, (i), por un lado, el HL es un sistema estructurado de signos interrelacionados, que denominamos términos y categorías, en el que el valor del significado de cada signo depende de su posición diferencial respecto de los demás, y que esta posición se define con el conjunto de relaciones que un signo del HL mantiene con los demás; y (ii), por otro lado, en un higraph, el valor del significado de cada signo se define aplicando el modelo semántico M= (D, μ) (sección 6.1).

El modelo semántico HL= (D, μ) se basa en definir el valor del significado de cada signo del tesauro comparándolo con los valores de los significados de los signos con los que está coordinado en el HL. En este sentido, se considera que el signo que se va a interpretar es el centro de una constelación, el punto donde convergen todos los demás signos del sistema lingüístico (el tesauro) relacionados con él. Esta constelación se va a representar, matemáticamente, como un conjunto de ternas, en el que cada terna representa la relación del signo interpretado con otro signo de la constelación de la forma siguiente: el primer componente es el signo que se interpreta, el segundo componente es la relación semántica que los une y el tercer componente es el signo coordinado.

$$\mu(\text{Signo}) = \{ (\text{Signo}, \text{TipoRelaciónSemántica}, \text{Signo\_coordinado}_n) \mid n \text{ es el número total de signos coordinados con Signo a interpretar}\}$$

Esta representación matemática es imprescindible para la manipulación automática, como veremos en los ejemplos de aplicación, sin embargo, no es comprensible para un usuario final. Por ello, proponemos que, además de una interpretación formal, el modelo semántico incluya una representación gráfica basada en los diagramas de Venn y en los grafos. Esta representación, además de complementar la representación formal matemática, es más intuitiva que las presentaciones normalizadas tradicionales de los tesauros, por lo que facilita la comprensión del contenido del tesauro a los usuarios, tanto a los que son expertos en el manejo de los tesauros como a los que no lo son.

Presentamos, en los siguientes epígrafes, de forma más detallada el modelo semántico HL separando cada uno de sus componentes: la interpretación de los términos y la interpretación de las categorías.

6.3.2.1. El cálculo del valor del significado de los términos

El modelo semántico de los H, M= (D, μ), calcula el significado de un glóbulo atómico aplicando una función de interpretación, μ, y el resultado es un conjunto de *valores semánticos* pertenecientes a dominio D. Formalmente:

$$\mu: T \rightarrow 2^D$$

Teniendo en cuenta la restricción de que

$$\text{si } X \neq Y \text{ entonces } \mu(X) \cap \mu(Y) = \varnothing,$$

En el HL los glóbulos atómicos son los términos del tesauro. Definimos el dominio semántico de valores de significado, D, como el conjunto de todas las posibles relaciones semánticas entre los signos del HL expresadas como ternas:

$$\mathbf{D = \{(x, rel, y) \mid x \in T, y \in S, rel \text{ es una relación semántica del HL}^{19}\}}$$
$$D = (T \times E \times T) \cup (T\times \{I^i\} \times \text{Categorías})$$

Formalmente D es un conjunto de ternas *término, etiqueta, término2*, que están formadas con todas las posibles combinaciones entre términos, relaciones semánticas y términos, y todas las posibles inclusiones de los términos en las Categorías. Como estamos considerando el caso de los tesauros, seleccionamos las relaciones semánticas estándar TG, TE, USE, USE PARA, TR. Las ternas *(término, I, categoría)* se refieren a la inclusión de los términos en las categorías: todas las posibles relaciones semánticas incluido en (representada con I) que pueden formarse entre los términos y las categorías de un tesauro dado. El nivel de inclusión de un término en una categoría lo indicamos con superíndice; así la secuencia de inclusiones: *término* $\subset$ *categoría1* $\subset$ *categoría2*; se representará con la terna *(término,* I^2*, categoría2)*[20].

[19] Definimos la relación I entre dos signos x e y de la forma: x I y sii x $\subset \sigma$(y); es decir, x está incluido en y si y sólo si x es un componente de y:. En ese caso utilizaremos la notación (x, I, y) y consideraremos que I es la relación inversa de σ.

[20] En general, consideramos que (x, I^i, y) sii x $\subset \sigma^i$ (y).

Por ejemplo, imaginemos un tesauro muy pequeño con tres términos (pájaro, árbol, ave), dos relaciones semánticas, TG y TR, y dos categorías (ANIMAL, VEGETAL). El dominio D de posibles valores semánticos sería:

D = {(pájaro, TG, pájaro), (pájaro, TG, árbol), (pájaro, TG, ave), (árbol, TG, árbol), (árbol, TG, pájaro), (árbol, TG, pájaro), (ave, TG, ave), (ave, TG, pájaro), (ave, TG, vegetal), (pájaro, TR, pájaro), (pájaro, TR, árbol), (pájaro, TR, ave), (árbol, TR, árbol), (árbol, TR, pájaro), (árbol, TR, pájaro), (ave, TR, ave), (ave, TR, pájaro), (ave, TR, vegetal), (pájaro I, animal), (ave, I, animal), (vegetal, I, animal), (pájaro, I, vegetal), (ave, I, vegetal), (vegetal, I, vegetal)}

En general, el tamaño del dominio D es el número de relaciones semánticas multiplicado por las posibles combinaciones los términos del HL tomados de dos en dos y, a este resultado, se le suman todas de inclusiones posibles de los términos en todas las categorías.

$$|D| = |E|\ C^2_{|T|} + |T| \times |\text{Categorías}|$$ [21]

En el tesauro ejemplo de las figuras 6.5 con sólo 15 términos, 6 categorías y 5 relaciones semánticas, el dominio semántico D tendrá un tamaño de $5 \times (15 \times 15) + 15 \times 6 = 1215$ "valores" posibles.

Ahora bien, de todo el dominio D de valores semánticos sólo algunos *tienen sentido* en una lengua[22], y estos son los que define la función μ. La función μ filtra los valores posibles, asignando a cada término sólo los que son correctos en una lengua. Formalmente, la función μ define la correspondencia entre cada término y un subconjunto de elementos de D, que son las relaciones en las que cada término participa con otros términos o categorías de un HL, y é*ste es el valor semántico del significado* de cada término:

$$\mu: T \rightarrow 2^D$$

μ(término) = {conjunto de relaciones semánticas en las que participa} =

= {término, relación, signo) | signo es un término o categoría del HL relacionado con relación con el término)}

[21] Se utiliza la notación matemática |D| para indicar el número total de elementos del conjunto D.
[22] Por ejemplo, es claro que la relación (Pájaro TG Pájaro) no tiene sentido en una lengua natural.

Por ejemplo, el término reptil del tesauro de la figura 6.5b, sólo participa en cuatro relaciones, que son las que se calculan aplicando la función μ:

μ(reptil) = {(reptil, TE, ofidio), (reptil, TE, tortuga), (reptil, TG, animal), (reptil, I, EUROPEO),
(reptil, I^2, ANIMAL SALVAJE)}

Esta interpretación se puede leer de la forma siguiente: "…el término *reptil* tiene como Término Específico al término *ofidio* y al término *tortuga*, como Término Genérico *animal* y es miembro de las categorías *EUROPEO* y a su vez, *ANIMAL SALVAJE*…", que es, exactamente, lo que debe entenderse cuando se lee la información asociada a *reptil* en el tesauro. La figura 6.6 visualiza esta interpretación[23].

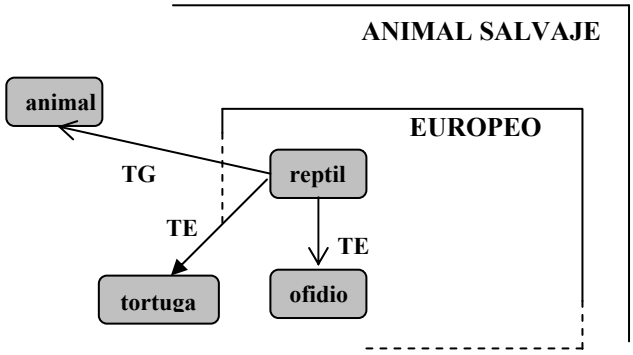


Figura 6.6. Interpretación gráfica del valor del significado del término 'reptil'

Además, y de acuerdo con el modelo semántico general de los higraphs, los valores semánticos de los términos son únicos (si X ≠Y entonces μ(X) ∩ μ(Y) = ∅). Esta restricción es importante en los tesauros porque significa, en términos del modelo de sistema de signos, que cada signo ocupa una *posición semántica única* en la estructura del tesauro. En el modelo semántico de HL que proponemos se cumple trivialmente, porque no es posible encontrar entre los significados de dos términos distintos dos tripletas iguales, como mínimo, el primer componente de la tripleta tiene que ser diferente:

μ(término 1) = {tripletas que comienzan por término 1: (término1, Relación, signo)}

μ(término 2) = {tripletas que comienzan por término 2: (término2, Relación, signo)}

Lo que garantiza que para dos términos cualquiera su posición es única:

μ(término 1) ∩ μ(término 2) = ∅

[23] De las tuplas (reptil, I, EUROPEO), (reptil, I^2, ANIMAL SALVAJE) puede inferirse que la categoría ANIMAL SALVAJE incluye a EUROPEO; sin embargo, no se puede saber qué otros términos y/o categorías incluyen ANIMAL SALVAJE y EUROPEO, ni las categorías a las que pertenecen los términos animal y ofidio. Por esta razón se dejan "abiertas" las categorías en la representación gráfica.

De esta forma, queda definida la interpretación semántica de los componentes atómicos de un HL de forma compatible con 1) el significado que un usuario espera obtener de un tesauro, 2) el principio de la solidaridad de los sistemas de las lenguas, y 3) el modelo semántico de los higraphs. Queda por definir el significado del resto de los componentes no atómicos del HL, el significado de las categorías.

6.3.2.2. El valor del significado de las categorías

El significado de las categorías, glóbulos no atómicos, se calcula incrementalmente a partir de los significados de los términos, componiendo la red con la unión y el producto cartesiano los significados de las categorías y términos que las constituyen. Formalmente:

$$\mu(\text{Categoría}) = \otimes_{(\forall i \in \{1..k\})} (\cup_{(\forall P \in \pi i(\text{Categoría}))} \mu(P))$$

Si la Categoría superior no se compone de facetas, su significado es la unión de los significados de sus componentes:

$$\mu(\text{Categoría}) = \cup_{(\forall P \in \sigma(\text{Categoría}))} \mu(P)$$

Por lo tanto, una categoría toma su significado de las categorías y términos que contiene. Por ejemplo, la categoría UNIVERSAL del HL ejemplo de la figura 6.5b tiene cinco componentes, la subcategoría ESPAÑOL y los términos *cerdo*, *chancho*, *guarro* y *puerco* (ver figuras 6.7 y 6.8). Su significado será la unión de los significados de estos componentes:

$$\mu(\text{UNIVERSAL}) = \mu(\text{ESPAÑOL}) \cup \mu(\text{cerdo}) \cup \mu(\text{chancho}) \cup \mu(\text{guarro}) \cup \mu(\text{puerco}))$$

Para calcular el significado de la categoría ESPAÑOL, aplicamos la composición por unión de sus componentes, en este caso sólo contiene el término *cerdo ibérico*:

μ(ESPAÑOL) = μ(cerdo ibérico) =
= {(cerdo ibérico, TG, cerdo), (cerdo ibérico, I, ESPAÑOL), (cerdo ibérico, I^2, UNIVERSAL), (cerdo ibérico, I^3, ANIMAL DOMÉSTICO)}

Teniendo en cuenta que:

μ(cerdo) = {(cerdo, TE, cerdo ibérico), (cerdo, USE PARA, chancho), (cerdo, TR, puerco), (cerdo, TR, guarro), (cerdo, TG, mamífero), (cerdo, TG, animal), (cerdo, I, UNIVERSAL), (cerdo, I^2, ANIMAL DOMÉSTICO) }

μ(chancho) = {(chancho, USE, cerdo), (chancho, I, UNIVERSAL) (chancho, I^2, ANIMAL DOMÉSTICO) }

μ(guarro) = {(guarro, TR, cerdo), (guarro, TR, puerco), (guarro, I, UNIVERSAL) (guarro, I^2, ANIMAL DOMÉSTICO)}

μ(puerco) = {(puerco, TR, cerdo), (puerco, TR, guarro), (puerco, I, UNIVERSAL) (puerco, I^2, ANIMAL DOMÉSTICO)}

El significado de la UNIVERSAL es:

μ(UNIVERSAL) = {(cerdo ibérico, TG, cerdo), (cerdo, TE, cerdo ibérico), (cerdo, TG, mamífero), (cerdo, TG, animal), (cerdo, USE PARA, chancho), (chancho, USE, cerdo), (cerdo, TR, puerco), (cerdo, TR, guarro), (guarro, TR, cerdo), (guarro, TR, puerco), (puerco, TR, cerdo), (puerco, TR, guarro), (cerdo ibérico, I, ESPAÑOL), (cerdo ibérico, I^2, UNIVERSAL), (cerdo, I, UNIVERSAL), (chancho, I, UNIVERSAL), (guarro, I, UNIVERSAL), (puerco, I, UNIVERSAL), (cerdo ibérico, I^3, ANIMAL DOMÉSTICO), (cerdo, I^2, ANIMAL DOMÉSTICO), (chancho, I^2, ANIMAL DOMÉSTICO), (guarro, I^2, ANIMAL DOMÉSTICO), (puerco, I^2, ANIMAL DOMÉSTICO)}

Una posible lectura sería: "… el cerdo ibérico es un cerdo, cerdo tiene un tipo llamado cerdo ibérico, cerdo es un mamífero, cerdo es un animal, cerdo tiene como sinónimo chancho, pero se prefiere utilizar cerdo en vez de chancho, cerdo, guarro y puerco son términos relacionados, cerdo ibérico es miembro de la categoría ESPAÑOL y, a su vez, de la categoría UNIVERSAL y, a su vez, de ANIMAL DOMÉSTICO; cerdo, puerco, guarro, chancho están incluidos de la categoría UNIVERSAL y también de la categoría ANIMAL DOMÉSTICO". La lectura de esta secuencia de "posiciones" nos permite (i) comprender el significado de esta categoría UNIVERSAL, (ii) nos permite *dibujar* el significado de la categoría UNIVERSAL (figura 6.7) y (iii) escribir la forma alfabética tradicional (figura 6.8).

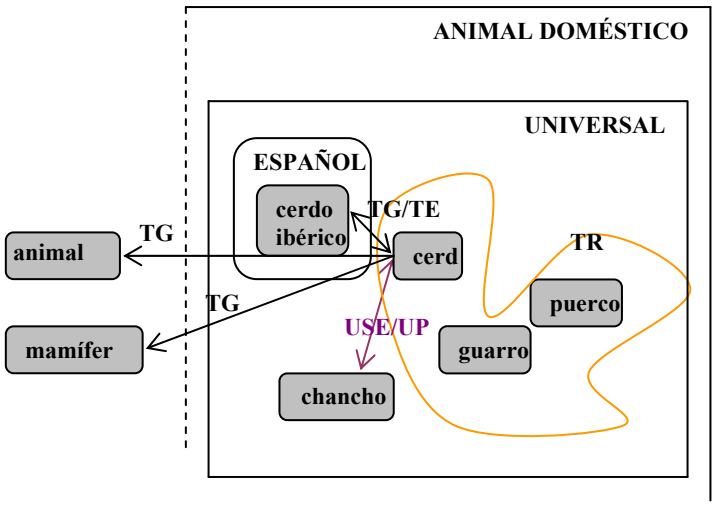


Figura 6.7. Interpretación gráfica del valor del significado de la categoría UNIVERSAL[24]

[24] La interpretación de UNIVERSAL incluye la interpretación completa de la subcategoría ESPAÑOL, por eso se dibuja como un conjunto "cerrado"; sin embargo, la categoría ANIMAL DOMÉSTICO, no pertenece a la interpretación de UNIVERSAL, surge de la interpretación de los términos que incluye, por lo que no puede dibujarse como un conjunto cerrado.

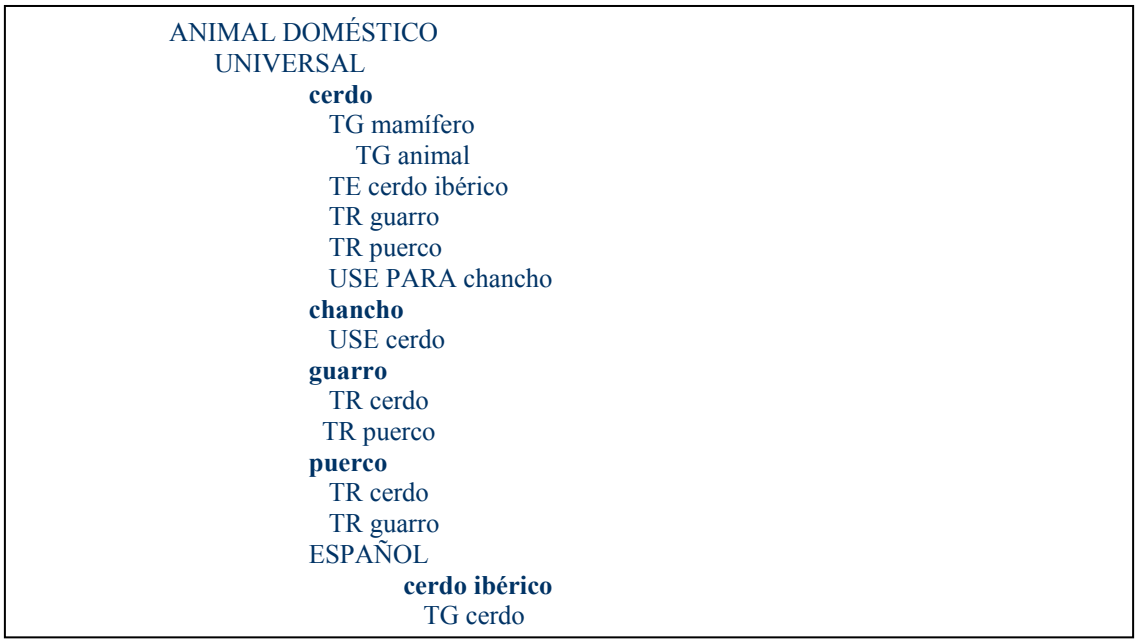


Figura 6.8. Categoría UNIVERSAL. Presentación alfabética

Para obtener una interpretación de todo el tesauro ejemplo, se comienza componiendo los significados de cada una de sus facetas ANIMAL SALVAJE y ANIMAL DOMÉSTICO con el producto cartesiano:

μ(TESAURO ejemplo) = μ(ANIMAL SALVAJE) $\otimes$ μ(ANIMAL DOMÉSTICO)

Los significados de cada faceta se obtienen de la misma forma que se ha hecho con la faceta UNIVERSAL. El resultado es un conjunto de posiciones, claramente separadas en dos componentes, la primera corresponde con las posiciones de la faceta ANIMAL SALVAJE, mientras que la segunda se corresponde con ANIMAL doméstico. Teniendo en cuenta que ANIMAL DOMESTICO sólo tiene un componente que es la categoría UNIVERSAL, μ(ANIMAL DOMÉSTICO) = μ(UNIVERSAL), aprovechamos los cálculos realizados anteriormente para escribir que:

μ(TESAURO ejemplo) = μ(ANIMAL SALVAJE) $\otimes$
{(cerdo ibérico, TG, cerdo), (cerdo, TE, cerdo ibérico), (cerdo, TG, mamífero), (cerdo, TG, animal), (cerdo, USE PARA, chancho), (chancho, USE, cerdo), (cerdo, TR, puerco), (cerdo, TR, guarro), (guarro, TR, cerdo), (guarro, TR, puerco), (puerco, TR, cerdo), (puerco, TR, guarro), (cerdo ibérico, I, ESPAÑOL), (cerdo ibérico, I^2, UNIVERSAL), (cerdo, I, UNIVERSAL), (chancho, I, UNIVERSAL), (guarro, I, UNIVERSAL), (puerco, I, UNIVERSAL), (cerdo ibérico, I^3, ANIMAL DOMÉSTICO), (cerdo, I^2, ANIMAL DOMÉSTICO), (chancho, I^2, ANIMAL DOMÉSTICO), (guarro, I^2, ANIMAL DOMÉSTICO), (puerco, I^2, ANIMAL DOMÉSTICO)}

Todo esto permite comprender que la interpretación de un HL, o categoría, dividido en facetas se realiza *interpretando de forma independiente y simultánea cada una de sus*

facetas. Esto es exactamente lo que se espera de un tesauro cuando está diseñado con facetas. Las facetas se utilizan cuando se desea separar el dominio de conocimiento que representa el tesauro en dos o más dimensiones. En este caso, para comprender el significado del tesauro se dirige al usuario a una lectura independiente de cada faceta, puesto que simultánea no parece posible, bien definiendo "dibujos" independientes por cada faceta si se presenta de forma gráfica, o bien con piezas de texto claramente separadas, si se presenta en forma sistemática (figura 6.9). La interpretación del dominio completo implica conocer el significado de todas sus facetas simultáneamente.

Figura 6.9. Tesauro facetado de Psicología del CSIC

En resumen, el HL es una aplicación del modelo de higraphs al dominio léxico que tiene el propósito de integrar, en un único modelo general, i) la expresividad del modelo matemático y visual de los higraphs, y ii) el principio de solidaridad sobre el valor del significado de los signos de los sistemas lingüísticos. El nuevo modelo se obtiene, a nivel sintáctico, estableciendo una correspondencia entre los elementos del modelo de higraphs y los elementos estándar de los tesauros monolingües. A nivel semántico, añadiendo, al modelo semántico del higraph, el valor que adquiere cada elemento en el tesauro. La interpretación del tesauro es la unión de sus componentes y la composición de cada una de sus partes.

El resultado es un modelo general, el HL, de alto nivel de abstracción para representar matemática y visualmente el contenido de cualquier tesauro. La aplicación de este modelo permite la explotación manual y automática del tesauro si se puede implementar un esquema de datos para el HL con modelos informáticos de representación de datos. Esto se puede probar, bien analizando la aplicabilidad, a la gestión de los HL, de los productos software que ya existen en el mercado para gestionar higraphs, o bien, encontrando una correspondencia entre el HL y un modelo de datos concreto, por

ejemplo, el modelo relacional de bases de datos, que permita explicar cómo construir y manipular HL.

6.4. Implementación del modelo HL

El objetivo en esta sección es crear una implementación de los HL conceptuales que sirva para comprobar y experimentar las posibilidades del modelo HL en la construcción de los tesauros. La forma más sencilla es utilizar alguna aplicación de carácter general para la gestión de higraphs, pero las que existen actualmente tienen una funcionalidad limitada; otra solución es crear el HL utilizando un modelo de implementación de datos con aplicaciones generales de creación y gestión. En el capítulo 5 hemos revisado los modelos de implementación de datos relacional y los modelos basados en XML aplicados a los tesauros y concluimos que, el modelo relacional es adecuado cuando el objetivo es una implementación eficaz, mientras que los modelos XML son adecuados cuando el objetivo es intercambiar, compartir y reutilizar el contenido de los HL. Para nuestros propósitos de experimentación del modelo HL, aplicamos el modelo relacional utilizando un SGBD para crear y manipular automáticamente la implementación del HL.

6.4.1. El uso de software de gestión de higraphs para la construcción y manipulación automática de los HL

Una de las formas de crear y explotar automáticamente un tesauro basado en HL es utilizando herramientas software para la gestión de higraphs. El HL es una aplicación de los higraph al dominio léxico, que puede ser construido y visualizado con cualquier herramienta para higraphs.

Una de las escasas herramientas para la gestión de higraphs del mercado es Structure 101. Esta herramienta está orientada a apoyar el diseño y desarrollo de aplicaciones informáticas mediante la visualización de los componentes y relaciones de dependencias entre los componentes que normalmente son muy complejas. Sin embargo, uno de los últimos desarrollos, la versión general Structure101g, amplía las posibilidades de aplicación de la herramienta a cualquier dominio de datos o de información. Los higraphs se definen utilizando un lenguaje XML propio de la herramienta. El único mecanismo de consulta de la información es la navegación (figura 6.10). Aunque esta visualización es clave para problemas de ingeniería inversa y para actividades que

impliquen explorar el higraph, para el propósito de esta investigación no es suficiente puesto que se necesita, además, poder crear, consultar y actualizar el HL.

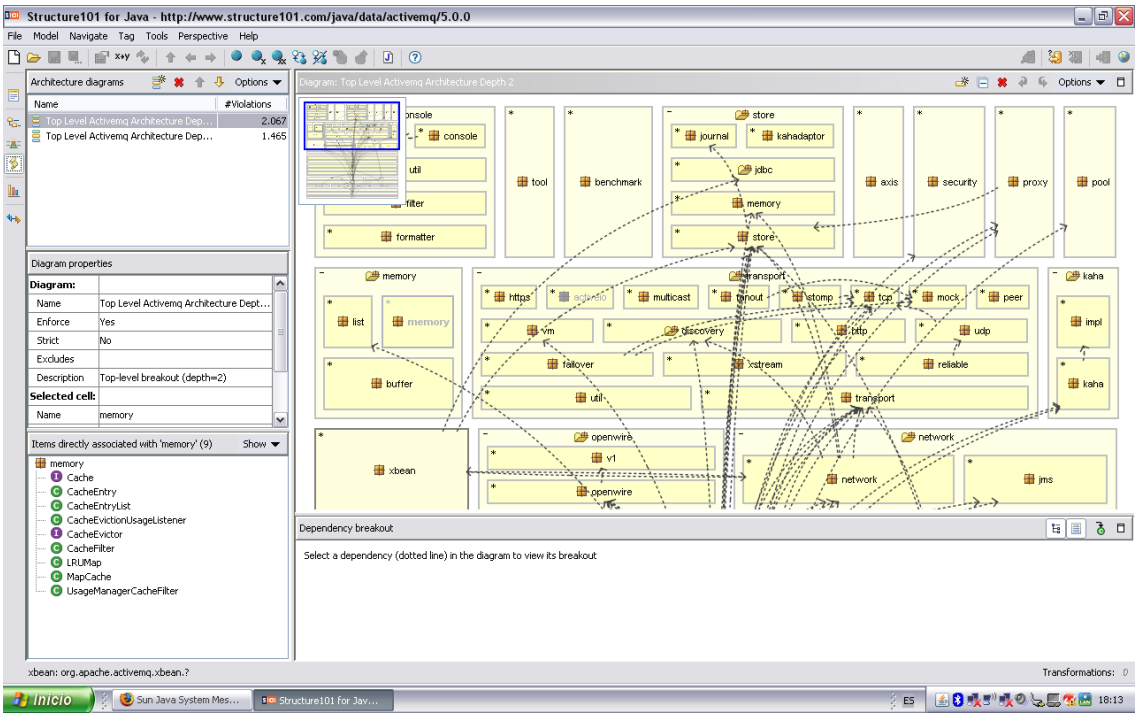


Figura 6.10. Visualización del Higraph de una aplicación software

6.4.2. El uso del modelo de datos relacional para la construcción y gestión automática de los HL

Una segunda aproximación a la implementación de HL es utilizar algún modelo de bases de datos que disponga de Sistemas de Gestión de Bases de Datos (SGBD) para construir y operar automáticamente sobre los componentes del HL. El modelo de implementación de datos relacional, además de disponer de SGBD, representa la información con relaciones matemáticas, es decir, con conjuntos de tuplas y esta *forma de organización es exactamente la misma que la propuesta para representar el significado de* un HL. En la sección anterior se han definido los valores del significado de los componentes de un HL mediante conjuntos de ternas (tuplas de tamaño 3), y el significado global del HL se obtiene componiendo los valores del significado de cada componente mediante las operaciones de unión y producto cartesiano. Estas operaciones son también operaciones del modelo relacional.

El modelo relacional tiene, además, otra ventaja para el propósito de este trabajo que es la posibilidad que ofrecen los actuales SGBDs de exportar e importar los datos en formatos estándares como texto plano con separadores, comas, espacios o tabuladores, de campo o texto etiquetado con XML. De esta forma, es posible la compartición de

datos y la interoperabilidad entre aplicaciones y bases de datos diferentes. Aunque estas ventajas hacen que el modelo relacional sea un modelo adecuado para la implementación del modelo HL, no es el único modelo aplicable. La selección del más apropiado requeriría un estudio detallado que considere cuál es el dominio de aplicación, los requisitos funcionales y de datos del HL, lo cual sobrepasa el objetivo que nos hemos marcado. Nuestro objetivo es demostrar la viabilidad de implementar el modelo conceptual de HL y obtener un esquema de datos para la construcción de los tesauros académicos de explotación. Para estos propósitos, el modelo relacional es razonablemente adecuado y es un modelo con el que estamos familiarizados por nuestra experiencia investigadora y docente anterior.

La idea, por lo tanto, es diseñar un esquema de datos relacional "general" para cualquier HL. Este esquema se utilizará para la construcción de cualquier tesauro, insertando, de forma incremental e inductiva, los términos, relaciones y categorías que se recojan de las fuentes, y que, en definitiva, son instancias –datos particulares- de los elementos del HL. En esta sección se propone y justifica *el esquema de datos relacional HL*. Este esquema HL es la clave para poder aplicar el método de construcción de tesauros que se presenta en el siguiente capítulo.

6.4.2.1. Diseño del HL relacional

El esquema de datos relacional HL está formado por un tipo de relación básica que denominamos *hl_macro*[25] y, de forma opcional, por otros tipos de relaciones auxiliares, cuyo propósito es completar o mejorar la eficiencia del HL incluyendo información auxiliar:

hl_macro (tipo de relación, signo1, signo2)

hl_macro representa los valores de μ(término) y μ(categoría). El primer componente del esquema de relación es el *tipo de relación semántica* cuyos valores, como hemos visto, son las relaciones semánticas entre términos (TG/TE, USE/USE PARA, TR y posiblemente otras específicas del dominio), pero también se incluyen como tipos de relaciones posibles el tipo de relación I (que significa 'miembro de' entre términos y categorías, e 'incluye' entre categorías y subcategorías), y la relación π, que significa 'es partición de', entre categorías y sus clases de equivalencia (que son las facetas del tesauro). Los dos componentes siguientes de *hl_macro* son los signos, que pueden ser

[25] Puesto que se corresponde con la macroestructura del tesauro.

término o categoría, relacionados con el *tipo de relación* marcado en el primer componente. Este esquema de relación recoge, por lo tanto, todos los valores semánticos de cada término y también los valores de las categorías, que excepto en el caso de las facetas, son la unión de los valores de sus subcategorías (ver, por ejemplo, tabla 6.1).

Este esquema de datos relacional recoge la estructura de un HL a partir de la representación de sus valores semánticos, es decir, la sintaxis y semántica respectivamente. Sin embargo, conviene indicar que existen dos diferencias con respecto al modelo semántico del HL. La primera, que no es relevante, es el cambio de orden entre el tipo de relación y el término. Se ha realizado porque facilitaba el tratamiento manual de los datos en las implementaciones experimentales iniciales[26]. La segunda es que el número de ternas de la relación *hl_macro* es menor que las de μ(UNIVERSAL). Esta diferencia sí que es relevante y se justifica por la necesidad de obtener un diseño relacional óptimo, que minimice la redundancia de los datos[27]. A modo de ejemplo, se reproducen los valores semánticos de la categoría UNIVERSAL para mostrar cómo se poblaría con un conjunto de instancias el esquema de relación *hl_macro*. Habitualmente las relaciones se representan con tablas porque son más sencillas de entender (tabla 6.1).

μ(UNIVERSAL) = {(cerdo ibérico, TG, cerdo), (cerdo, TE, cerdo ibérico), (cerdo, TG, mamífero), (cerdo, TG, animal), (cerdo, USE PARA, chancho), (chancho, USE, cerdo), (cerdo, TR, puerco), (cerdo, TR, guarro), (guarro, TR, cerdo), (guarro, TR, puerco), (puerco, TR, cerdo), (puerco, TR, guarro), (cerdo ibérico, I, ESPAÑOL), (cerdo ibérico, I^2, UNIVERSAL), (cerdo, I, UNIVERSAL), (chancho, I, UNIVERSAL), (guarro, I, UNIVERSAL), (puerco, I, UNIVERSAL), (cerdo ibérico, I^3, ANIMAL DOMÉSTICO), (cerdo, I^2, ANIMAL DOMÉSTICO), (chancho, I^2, ANIMAL DOMÉSTICO), (guarro, I^2, ANIMAL DOMÉSTICO), (puerco, I^2, ANIMAL DOMÉSTICO)}

tipo	**signo1**	**signo2**
I	cerdo ibérico	ESPAÑOL
I	cerdo	UNIVERSAL
I	chancho	UNIVERSAL
I	guarro	UNIVERSAL
I	puerco	UNIVERSAL
I	UNIVERSAL	ANIMAL DOMÉSTICO
I	ESPAÑOL	UNIVERSAL

[26] Este cambio de orden se ha realizado porque facilitaba la lectura en las tablas de datos.

[27] La redundancia de los datos genera, a lo largo de la vida de una base de datos problemas de inconsistencia, puesto que cuantas más veces esté repetido un dato, más probable será olvidarse de actualizarlo en alguno de los lugares donde aparece.

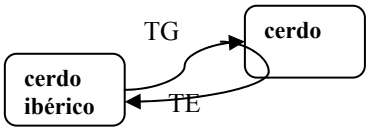
I	TESAURO ejemplo	ANIMAL DOMÉSTICO
TG	mamífero	animal
TG	reptil	animal
TG	lobo	mamífero
TG	cerdo	mamífero
TG	bóvido	mamífero
TG	muflón	bóvido
TG	ofidio	reptil
TG	tortuga	reptil
TG	cerdo ibérico	cerdo
USE	irracional	animal
USE	bicho	animal
USE	chancho	guarro
TR	cerdo	guarro
TR	puerco	guarro

Tabla 6.1. Tabla con los valores de la relación hl_macro correspondientes a la categoría UNIVERSAL. Se puede observar que contiene menos ternas que μ(UNIVERSAL)

Los criterios básicos para el diseño de esquemas óptimos de bases de datos relacionales son, siguiendo a Ullman (1988:1) la exhaustividad de la representación, ya que no debe perderse información, 2) la coherencia de los datos, ya que no deben existir inconsistencias, y 3) mínima redundancia, puesto que debe existir la mínima duplicación de datos posible. La aplicación de estos criterios en los HL ha supuesto resolver cuatro cuestiones: (i) la representación de las relaciones semánticas binarias y asimétricas; (ii) la representación de las relaciones simétricas entre varios elementos, los hiperarcos; (iii) la representación de las relaciones de orden, asimétricas y transitivas que generan jerarquías; y (iv) la redundancia que genera las relaciones transitivas.

La primera cuestión que hay que resolver es la representación de las relaciones semánticas binarias y asimétricas que, en el modelo relacional, conlleva un problema de duplicación de datos. Estas relaciones tienen dos sentidos diferentes con significados inversos. Para que el significado de una relación asimétrica sea completo y coherente es necesario que se expliciten los dos sentidos. En μ(UNIVERSAL), por ejemplo, tenemos dos ternas por cada relación asimétrica que son (cerdo ibérico, TG, cerdo) y (cerdo, TE, cerdo ibérico) (figura 6.11)[28].

[28] Sin embargo, en el caso de las relaciones simétricas sólo es necesaria una terna por cada relación simétrica. Ver (cerdo, TR, guarro).



Relación 1: (TG, cerdo ibérico, cerdo)
Relación 2: (TE, cerdo, cerdo ibérico)

Figura 6.11. Repetición de datos en la relación binaria asimétrica

Esta circunstancia es un inconveniente desde el punto de vista de la gestión y mantenimiento de la base de datos relacional, porque implica insertar dos veces los mismos términos, es decir, duplicar los datos. Esta redundancia se puede evitar si se establece y documenta un convenio para seleccionar siempre uno de los sentidos de las relaciones asimétricas: la posición de los argumentos de la relación marca el papel que juegan estos argumentos en la relación. En el ejemplo de la tabla 6.1, se ha seleccionado la etiqueta TG para representar las relaciones TG y TE, lo que significa que los valores de la segunda columna de la tabla son siempre los Términos Genéricos, mientras que los de la tercera columna son los Términos Específicos. Cuando se necesita acceder al término genérico se busca la segunda columna de una fila con la etiqueta TG. En el caso inverso, el resultado estará en la tercera columna. La figura 6.12 muestra, con un ejemplo, cómo se obtendrían los tipos de relaciones inversas en *hl_macro* utilizando el lenguaje SQL.

Término genérico de *cerdo*:
SELECT signo1
FROM *hl_macro*
WHERE tipo relación LIKE “TG” AND signo LIKE “cerdo”

Término específico de *cerdo*:
SELECT signo1
FROM *hl_macro*
WHERE tipo relación LIKE “TG” AND signo LIKE “cerdo”

Figura 6.12. Acceso al término genérico o específico de cerdo utilizando SQL

En consecuencia, en la representación relacional del HL, tomamos por convenio que las relaciones jerárquicas se representan sólo con una etiqueta, por ejemplo TG; y las relaciones de sinonimia o variante léxica se representan con la etiqueta USE.

La segunda cuestión de diseño se refiere al mecanismo de representación de los hiperacos. Un hiperarco representa las relaciones simétricas entre más de dos elementos,

como es el caso de la relación TR. En el modelo del HL los hiperarcos se traducen a un conjunto de arcos binarios que relacionen todos los nodos del hiperarco entre sí. Esta es la solución que se adoptó en el modelo semántico HL con la relación TR: se definen tantas ternas como combinaciones existen entre un término y sus términos relacionados. –ver, por ejemplo, el caso de *cerdo*, *guarro* y *puerco*. Esta solución, sin embargo, supone una repetición de términos poco deseable que *debería poder optimizarse, reducir el número de ternas.*

El mecanismo que hemos diseñado para obtener una representación mínima de ternas para un hiperarco lo denominamos *modelo en estrella* (figura 6.13) y consiste en seleccionar uno de los nodos del hiperarco como *pivote* o centro[29]. El resto de los nodos del hiperarco deben relacionarse sólo con el pivote y, gráficamente, formarían una estrella[30]. Este esquema se traduce, en el modelo relacional, en tuplas en las que el pivote está siempre en la misma posición, por ejemplo la primera, y las puntas están en la otra posición. En el caso de la figura 6.13, la estrella se representaría con las dos ternas (TR, cerdo, puerco), (TR, cerdo, guarro). Se evitan por lo tanto las tuplas redundantes: (TR, cerdo, puerco), (TR, cerdo, guarro), (TR, guarro, cerdo), (TR, guarro, puerco), (TR, puerco, cerdo), (TR, puerco, guarro).

Respecto a las consultas, para conocer todos los términos de un hiperarco TR[31], se propone el siguiente procedimiento con tres pasos:

- paso 1: se seleccionan las filas de la tabla *hl_macro* que tienen como primera columna TR y el término buscado, por ejemplo *guarro*, en cualquiera de las otras columnas;
- paso 2: si el término, por ejemplo 'guarro', está en la tercera columna, se busca el valor de la segunda que es el pivote, 'cerdo' en este caso, y es una de las soluciones. Si el término buscado estuviera en la segunda columna, es directamente el pivote y se pasa al paso 3; y

[29] Por ejemplo, en la figura 6.13 se selecciona 'cerdo' como pivote porque en el DRAE (2001) se considera este término como forma más general.

[30] Esta implementación es una variante de la propuesta del el w3c para definir relaciones N-arias entre participantes con el mismo grado de protagonismo y con papeles diferentes en la relación (ver Use Case 3 en Noy y Rector, (2006:8)). El caso de las relaciones semánticas asociativas, TR, es más simple porque todos los participantes tienen el mismo papel en la relación. Por eso, en vez de crear un elemento nuevo para representar la relación, como se hace en la propuesta del w3c, nosotros proponemos que sea uno de los términos el que sirva de representante de la relación, y le denominamos "pivote". Esta opción simplifica la representación de la relación y la construcción incremental e inductiva de estas relaciones.

[31] También para consultar si un par de términos del hiperarco están relacionados, por ejemplo, para consultar cuál es la relación entre 'puerco' y 'guarro'.

- paso 3: el resto de las soluciones se buscan a partir del término pivote. Son todos los valores de las terceras columnas tales que en la primera columna tengan la relación TR y en segunda columna tenga el término pivote.

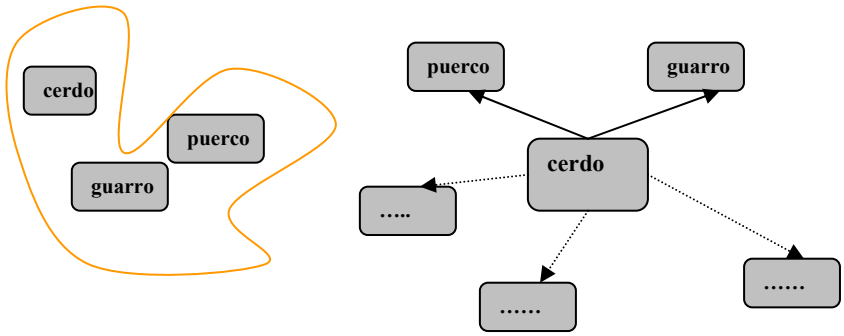
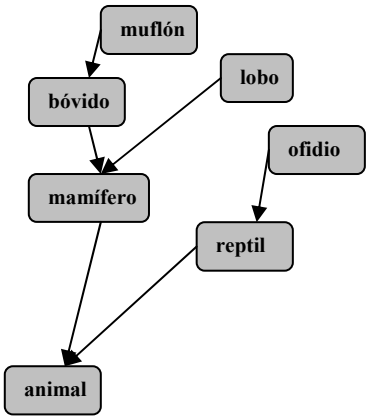


Figura 6.13. Modelo en estrella para resolver la representación de los hiperarcos en el modelo relacional

La tercera cuestión de diseño se refiere a la representación en el esquema *hl_macro* de las jerarquías generadas por relaciones asimétricas y transitivas como la relación TG (figura 6.14). El modelo relacional es un modelo 'plano' en el sentido de que los datos se representan mediante estructuras, las relaciones, que son colecciones de valores en un único nivel. Las estructuras jerárquicas son colecciones de valores en varios niveles, y el recorrido de todos los niveles se realiza utilizando consultas recursivas. Hasta hace pocos años, esto era un problema en el modelo relacional porque la especificación estándar del lenguaje de consulta SQL no disponía de la capacidad de definir consultas recursivas. Se necesitaban lenguajes de programación de propósito general para poder manipular las jerarquías, lo que complicaba el diseño y gestión de las bases de datos. Recientemente se ha extendido el estándar SQL con mecanismos de recursividad que permiten el tratamiento directo de datos estructurados en jerarquías (ISO/IEC 9075-14, 2008), lo que ha facilita la representación y gestión de las relaciones jerárquicas (como TG).

La representación de una jerarquía en el esquema *hl_macro* es exactamente la misma que en el modelo HL: un conjunto de tuplas, cada una con la especificación del tipo de relación, el "hijo" y el "padre" (figura 6.14).



(TG, animal, **mamífero**)
(TG, **mamífero**, **bóvido**)
(TG, **bóvido**, muflón)
(TG**,** **mamífero**, lobo)
(TG, animal, reptil)
(TG, reptil, ofidio)

Figura 6.14. Jerarquía de relaciones TG (izq.) y su representación relacional (dcha)

El procedimiento para acceder a todos los nodos de la jerarquía es ir conectando en una misma relación los nodos hijo de una terna, que están en la tercera posición, con los nodos padre, que están en la segunda posición, de otra terna[32]. En SQL esta condición se expresa como:

(PRIOR rels.name = rels.name AND PRIOR signo2 = signo1)

La cuarta cuestión se refiere también a las relaciones asimétricas y transitivas, relaciones de orden, pero, concretamente, a la redundancia que genera la transitividad. En la relación I, los términos se repiten porque pertenecen a varias categorías que son subcategorías entre sí. Por ejemplo, el término *cerdo ibérico* pertenece a la categoría ESPAÑOL[33], la categoría ESPAÑOL está incluida en la categoría UNIVERSAL[34], y *cerdo ibérico*, por lo tanto, pertenece a la categoría UNIVERSAL.

tipo	**Signo1**	**Signo2**
I	cerdo ibérico	ESPAÑOL
I	cerdo ibérico	UNIVERSAL
I	ESPAÑOL	UNIVERSAL

Esta redundancia puede resolverse explicitando las inclusiones de unas categorías en otras, que en el modelo semántico aparecen implícitas, e incluyendo los términos únicamente en las categorías más específicas. De esta forma, la segunda de las tuplas del ejemplo anterior, (I, cerdo ibérico, UNIVERSAL), no formaría parte de la relación *hl_macro*.

En definitiva, el esquema *hl_macro(tipo_relación, signo1, signo2)* es una representación relacional del modelo HL en la que tipo_relación tiene valores que son

[32] Se ha marcado en la figura 6.11 con colores las conexiones.
[33] Porque es un animal propio de España.
[34] Porque es un animal extendido en todos los países.

los tipos de relaciones semánticas de un tesauro. Las relaciones semánticas para la organización de los datos se clasifican en tres tipos:

1) Relaciones asimétricas. Sólo se representa uno de los tipos de relaciones, la relación inversa se calcula mediante "consultas inversas". Por ejemplo la relación USE
2) Relaciones simétricas. Muchos a muchos (hiperárcos) deben conformar un modelo en estrella, con un elemento pivote con el que se relacionan todos los demás. Por ejemplo la relación TR
3) Relaciones de orden (asimétricas y transitivas). Se representan las relaciones directas, la transitividad se obtiene mediante consultas (recursivas). Por ejemplo las relaciones TG y sigma.

Para mantener la consistencia de la base de datos se añade a su esquema una relación más, *hl_micro(signo, tipo),* con todos los signos del HL y su tipo, término o categoría. Esta segunda relación evita que existan errores tipográficos en los signos porque están definidos sólo una vez en *hl_micro*; en *hl_macro* se utilizan los identificadores de los términos y, antes de la inserción, el SGBD comprueba automáticamente que están previamente definidos en hl_micro (figura 6.15).

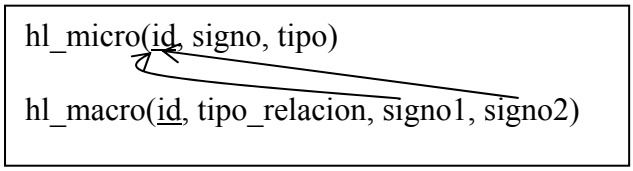


Figura 6.15. Esquema de datos relacional básico del HL

Este esquema básico puede ampliarse si se necesita incluir más información en el tesauro. Por ejemplo, las notas de ámbito que refinan y desambiguan el significado de los términos se incluyen como un atributo en la tabla *hl_micro*.

6.4.2.2. Ejemplo

El tesauro europeo del repositorio ARIADNE-SILO[35], el tesauro LRE[36], es un ejemplo interesante de aplicación del modelo HL porque dispone de una interfaz gráfica de navegación que representa un HL en el que la única relación semántica es I, la inclusión de los términos en categorías (figura 6.16).

[35] SILO (Search & Index Learning Objects) es un repositorio europeo de objetos de aprendizaje en 12 lenguas, de acceso libre, previo registro. Disponible en: http://ariadne.cs.kuleuven.be/silo2006/Welcome.do;jsessionid=3941026FFED2896392EA09CFD87A2510

[36] El 'Learning Resource European thesaurus' se describe en: http://lre.eun.org/node/6

Este tesauro es también uno de los paradigmas de tesauro de explotación de referencia de un repositorio de recursos educativos porque constituye una de las herramientas de acceso y exploración del repositorio ARIADNE-SILO. Tiene dos facetas principales: 1) *Exact, Natural and Engineering Sciences y* 2) *Human and Social Sciences*, que se van refinando en jerarquías de 3 niveles de profundidad, y no existen las relaciones semánticas TG/TE, USE/USE FOR, TR.

Aplicando el modelo de implementación de datos relacional HL a una de las facetas del tesauro, "Human and Social Sciences" (figura 6.17), se obtiene la base de datos mostrada en las tablas 6.3 y 6.4.

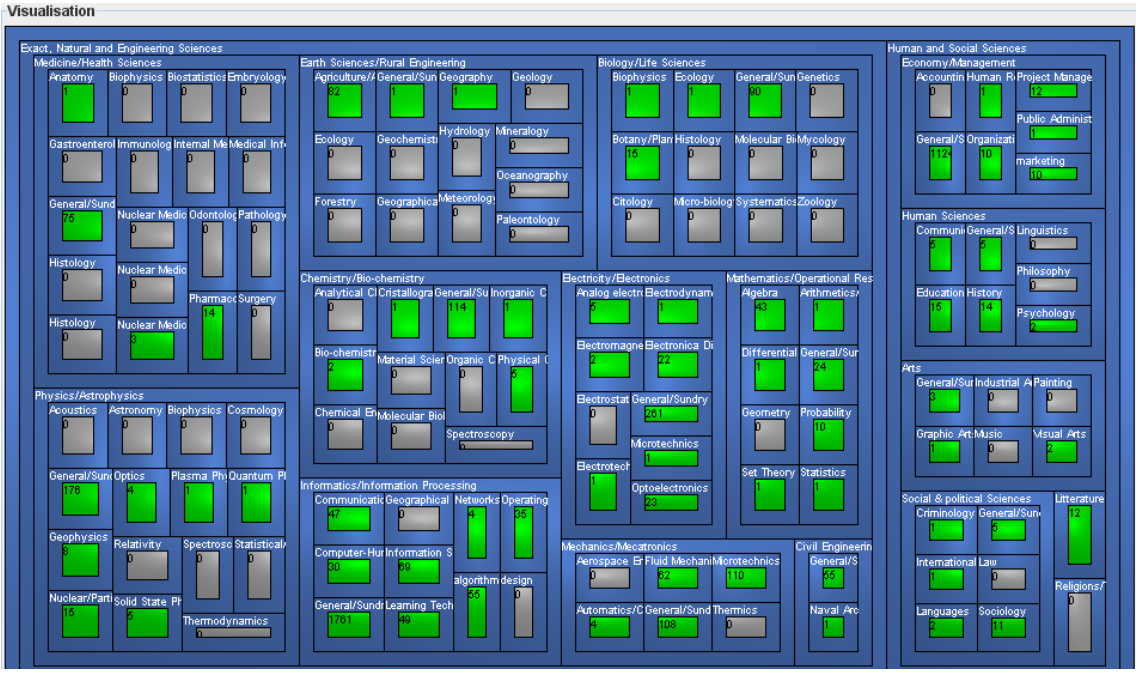


Figura 6.16. Interfaz gráfica del tesauro del repositorio SILO[37]

En este tesauro, el significado de los términos y categorías depende sólo de las relaciones jerárquicas σ (y la relación inversa I) y de la relación de partición π. Por ejemplo, el significado del término *marketing* (figura 6.17 esquina superior derecha) es su valor semántico μ(marketing), que, a su vez, es el conjunto de relaciones que mantiene con las categorías del HL. :

μ(marketing) = {(I, marketing, Economy/Management), (I, marketing, Human and Social Sciences)}

[37] Fuente: http://ariadne.cs.kuleuven.be/silo2006/visualbrowse.jsp

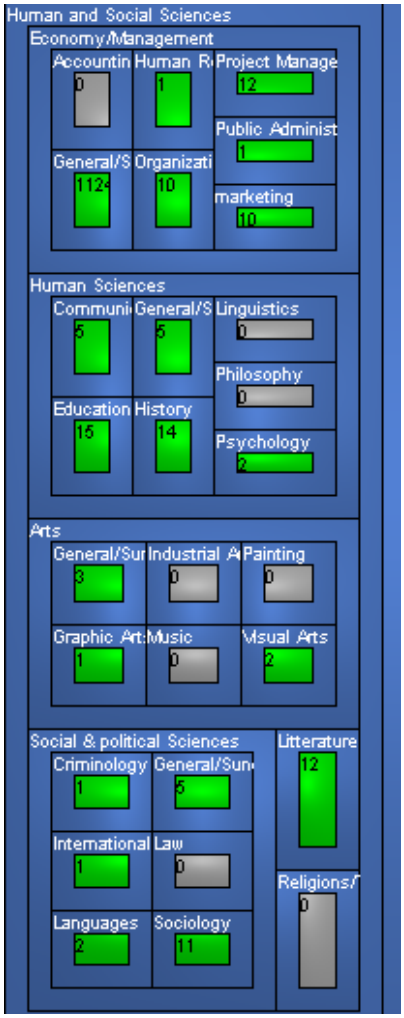


Figura 6.17. Faceta "Human and Social Sciences"

signo	tipo
Accounting	term
Human Resources	term
Proyect Management	term
Public Administration	term
General/Sundry	term
Organization	term
Marketing	term
Comunication	term
Linguistics	term
Education	term
Philosophy	term
Psicology	term
History	term
Industrial Art	term
Painting	term
Music	term
Visual Arts	term
Graphic Arts	term
Criminology	term
Internacional relations	term
Law	term
Languages	term
Sociology	term
Religión/Theology	term
Litterature	term
Human and Social Sciences	category

Economy/Management	category
Human Sciences	category
Arts	category
Social & political Sciences	category

Tabla 6.3. Relación hl_micro

tipo	**signo1**	**signo2**
I	Human and Social Sciences	Economy/Management
I	Human and Social Sciences	Human Sciences
I	Human and Social Sciences	Arts
I	Human and Social Sciences	Social & political Sciences
I	Economy/Management	Accounting
I	Economy/Management	Human Resources
I	Economy/Management	Proyect Management
I	Economy/Management	Public Administration
I	Economy/Management	General/Sundry
I	Economy/Management	Organization
I	Economy/Management	Marketing
I	Human Sciences	Comunication
I	Human Sciences	Linguistics
I	Human Sciences	Education
I	Human Sciences	Philosophy
I	Human Sciences	Psicology
I	Human Sciences	History
I	Arts	Industrial Art
I	Arts	Painting
I	Arts	Music
I	Arts	Visual Arts
I	Arts	Graphic Arts
I	Social & political Sciences	Criminology
I	Social & political Sciences	Internacional relations
I	Social & political Sciences	Law
I	Social & political Sciences	Languages
I	Social & political Sciences	Sociology
π	Human and Social Sciences	Economy/Management
π	Human and Social Sciences	Human Sciences
π	Human and Social Sciences	Arts
π	Human and Social Sciences	Social & political Sciences

Tabla 6.4. Relación hl_macro

En definitiva, este ejemplo muestra cómo se crearía un tesauro con el esquema de datos relacional HL y cómo los diagramas gráficos que se obtendrían sirven para entender mejor el ámbito del tesauro y, en el caso de tesauros de explotación, para guiar al usuario en la búsqueda de la información o recursos que necesita.

6.5. Resumen y conclusiones del capítulo

El modelo de higraph léxico (HL) que hemos presentado es una aplicación del modelo de higraphs al dominio léxico. Para definir la sintaxis se establece una correspondencia entre los elementos de un higraph y los elementos estándares de un tesauro. En consecuencia, la estructura del HL es una particularización de las estructuras de los higraphs. La semántica, sin embargo, es una extensión de la semántica de los higraphs.

Esta extensión tiene el propósito de incluir el concepto de valor del significado de los términos de un tesauro, entendido éste como sistema lingüístico sujeto al principio de solidaridad. Tal y como hemos convenido, valor semántico de un término es el conjunto de relaciones en las que participa dentro del HL. El valor semántico de una categoría es la composición incremental, aplicando el modelo semántico de los higraphs, de los significados de los términos y categorías que contiene.

En estos momentos no existe un software de carácter general para crear y gestionar los higraph ni los HL; en consecuencia, es necesario construir, con un modelo de implementación de datos adecuado, un esquema de datos general para crear los HL. Este esquema servirá para organizar y gestionar el contenido de los diferentes tesauros en un entorno informático. Para este propósito, el modelo relacional es adecuado porque (i) las relaciones, que son las estructuras de datos básicas del modelo, representan de forma directa la interpretación del esquema conceptual de un higraph, y, además, (ii) es razonablemente sencillo construir y gestionar con los SGBD disponibles.

El modelo HL es un tipo de higraphs para representar el conocimiento léxico de los tesauros (figura 6.18). Los modelos de tesauro, creados con las metodologías tradicionales, mediante un análisis conceptual del dominio o de las fuentes de términos, se pueden representar con un HL. Desde este punto de vista, se puede definir el HL como un metamodelo que aporta (i) una arquitectura general para todos los tesauros, que, además, puede visualizarse, (ii) una implementación eficiente, (iii) escalabilidad, y (iv) mayor flexibilidad en la gestión del contenido los tesauros. Se prueba, así, una parte de la hipótesis de este trabajo de investigación: es posible encontrar un modelo general para expresar y sistematizar el conocimiento contenido en cualquier tesauro, con independencia del ámbito, propósito y tipo de usuario.

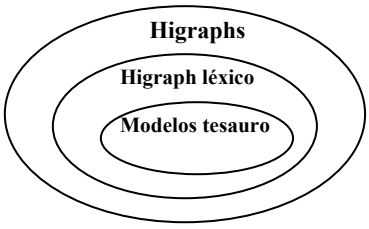


Figura 6.18. Niveles de abstracción de los modelos

Capítulo 7

Una metodología para la construcción inductiva de tesauros académicos de explotación

"God does not care about our mathematical difficulties. He integrates empirically"
A.Einstein

En este capítulo se demuestra la segunda hipótesis de este trabajo de tesis: es posible encontrar una estrategia para construir y mantener los tesauros académicos de explotación que reproduzca la estructura conceptual y el lenguaje de especialidad de los docentes e investigadores con un coste menor que si se utilizan las metodologías establecidas de forma general para la construcción de tesauros. Para ello se propone un método inductivo que construye los tesauros de sentido específico a genérico, en un proceso continuo e integrado en la actividad de creación, catalogación y uso de los recursos didácticos digitalizados hecha por los expertos en la materia: profesores, investigadores y estudiantes. Este método extrae estructuras terminológicas en semántica libre[1] de los recursos didácticos y las incorpora a un HL. Cuando hablamos de estructuras terminológicas en semántica libre nos referimos a pequeñas redes de términos con relaciones semánticas, una o varias simultáneamente, que no están previamente establecidas, que están inmersas en el contenido y/o meta-contenido de materiales educativos y que son propuestas por uno o varios especialistas de esa comunidad de forma libre, es decir, por medio de una elección libre, lo que no implica que sean originales o únicas.

El resultado es tesauros académicos de explotación, que describen y organizan el conocimiento y los contenidos o los recursos didácticos específicos que han servido de fuente de estructuras terminológicas y que puede ir incrementalmente creciendo y adaptándose a las nuevas fuentes que los autores vayan incorporando. En este capítulo presentaremos nuestra propuesta para la construcción de este tipo de tesauros, pero antes revisaremos los métodos generales para la construcción de tesauros que han servido de base al método nuevo.

[1] Elegimos esta denominación por analogía con 'sintaxis libre' que supone estructuras no consolidadas en la lengua como formas de cita (Lyons, 1977).

7.1. Métodos de construcción de tesauros

El objetivo de un tesauro es reproducir la estructura conceptual de un dominio de conocimiento utilizando términos relacionados semánticamente, preferentemente con relaciones de generalización/especialización, equivalencia y asociativas. El proceso de construcción de un tesauro puede ser automático, manual o inteligente y semi-automático pero en cualquiera de estos casos, se dispone de un conjunto de principios y recomendaciones bien establecidos, definidos en los estándares para la construcción de tesauros mono o multilingües (ver capítulo 4) para definir: 1) la presentación y marcado del contenido del tesauro; y 2), el proceso de construcción.

7.1.1. El proceso de construcción

La construcción es, habitualmente, un proceso incremental, formado por una secuencia, iterativa, de operaciones cuyo orden depende de la aproximación metodológica, deductiva o inductiva, que se aplica (Aitchinson et al., 2000, Soergel, 2004). La figura 7.1 muestra el proceso. Las operaciones están representadas con un recuadro y los productos resultantes sin recuadro. No se han incluido las iteraciones que, siempre, se realizan entre cada par de operaciones para comprobar si los resultados de una operación son coherentes con los de la operación anterior; sin embargo, sí que se han incluido las iteraciones que se realizan para confirmar o reajustar el esquema conceptual, y la que se realiza al final del proceso cuando se evalúa el tesauro durante el uso que, de no ser satisfactorio, puede obligar a revisar todo el proceso hasta que los resultados sean coherentes. De forma más detallada estas operaciones son:

1) Análisis del dominio: son los preliminares donde se establece el objetivo y naturaleza del tesauro y determinará el resto de las operaciones. El resultado es, normalmente, un documento que describe el plan de la obra, y que sirve, habitualmente, para documentar el tesauro cuando se publica (ver, por ejemplo, documento del BCD del CES para la actualización del tesauro OIT[2] (Lorite, 2004), o los preliminares de los tesauros del CINDOC[3]). En el plan se definen los objetivos y naturaleza del tesauro, concretamente:

[2] Tesauro de la Organización Internacional del Trabajo (OIT), es un tesauro multilingüe del ámbito laboral. Versión española disponible en: http://www.ilo.org/public/libdoc/ILO-Thesaurus/spanish/

[3] El CINDOC es, actualmente, el Instituto de Estudios Documentales sobre Ciencia y Tecnología (IEDCyT). Los tesauros están disponibles en: http://thes.cindoc.csic.es/

i) las características de presentación del contenido en los diferentes soportes, papel o electrónico, en los que se necesite reproducir el tesauro;

ii) los tipos de uso, que es lo que constituye, realmente, el objetivo del tesauro; por ejemplo, el aprendizaje, la indexación, clasificación y recuperación de información o recursos, ayuda a la escritura;

iii) El tipo de usuarios al que va dirigido el tesauro, expertos del dominio o público en general. Aquí conviene apuntar que, normalmente, los usuarios de un tesauro no son los creadores del mismo, pero en el entorno académico es habitual el caso contrario en el que el profesor es el creador y el usuario. Los alumnos, por su parte, son usuarios y en ciertos casos también participan, dentro de actividades de carácter didáctico o investigador, en la construcción del tesauro;

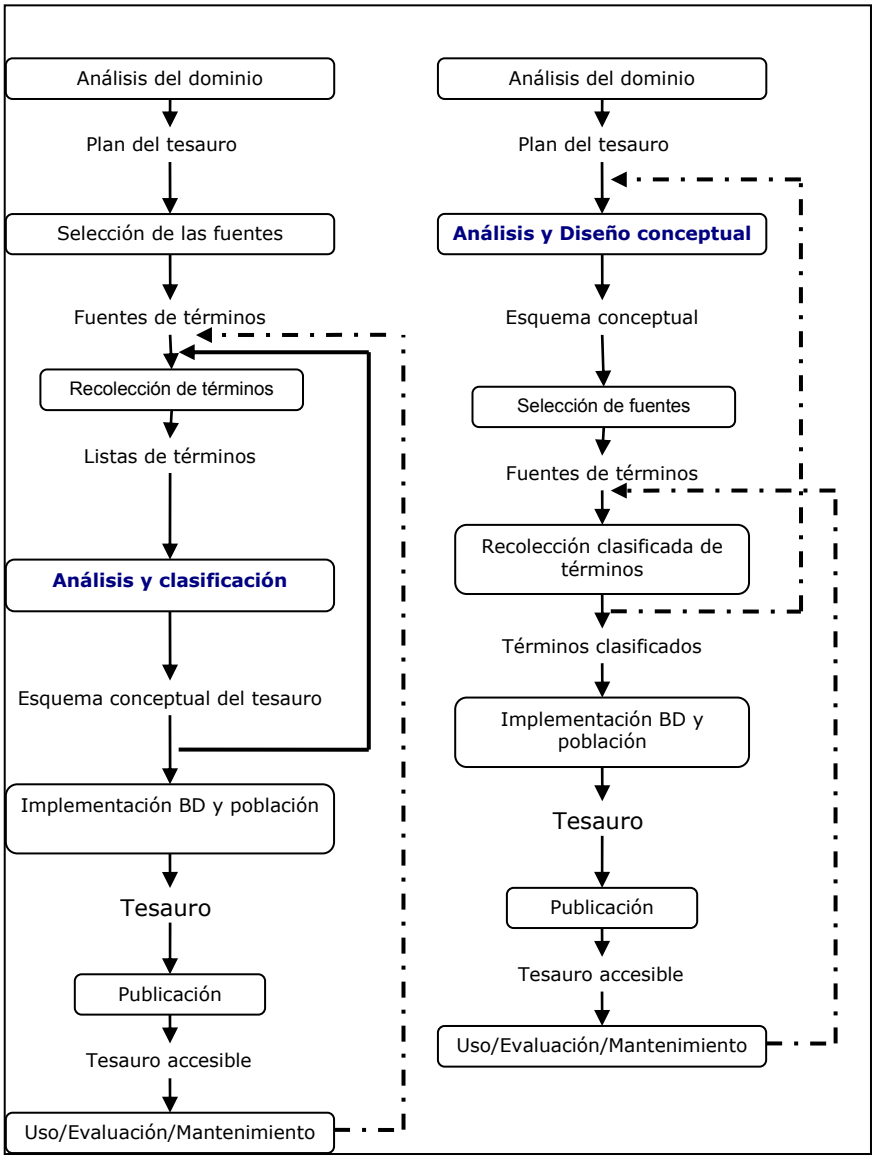


Figura 7.1. Proceso de construcción de tesauros inductivo (izquierda), deductivo (derecha)

iv) La cobertura, que es el tamaño del tesauro, tanto desde el aspecto extensivo, número de términos, como intensivo, número de campos de información en cada término. Está delimitado por el dominio de conocimiento o por el conjunto de objetos que tienen que describirse y clasificarse, por el tipo de uso y por el tipo de usuario;

v) Los requisitos técnicos, que dependen fundamentalmente de si el tesauro se plantea como:
   - Un sistema independiente, y/o
   - Estará integrado en algún sistema de información, como herramienta de acceso a la información y a los recursos que gestione el sistema. En este caso, el tesauro debe considerarse como un "índice semántico";

vi) El tipo de consultas y otras operaciones automáticas que se van a realizar conforme a los tipos de uso. En esta fase se debe decidir, puesto que son contrapuestos, el grado de:
   - exhaustividad, lo que implica definir:
     - un control sobre la forma de la palabra,
     - control sobre los sinónimos y cuasi-sinónimos,
     - añadir mecanismos de búsqueda sobre las partes de las palabras,
     - mecanismos de búsqueda que vayan de términos específicos a más generales, y
     - añadir términos cercanos semánticamente en la red; y
   - precisión, lo que supone definir:
     - mayor cantidad de términos específicos,
     - la coordinación de términos, en el indexado o en las búsquedas,
     - homógrafos y notas de ámbito,
     - mecanismos de búsqueda desde los términos más genéricos a los términos específicos, y
     - una indicación de términos adyacentes.

Asimismo, en el plan se incluye la planificación del trabajo y la gestión del personal, los recursos, el tiempo, así como los métodos de evaluación y, finalmente, el sistema de mantenimiento y actualización del tesauro.

2) Selección de fuentes. Consiste en definir las obras y documentos que servirán para extraer los términos del futuro tesauro. El contenido de las fuentes depende del ámbito,

la cobertura y el tipo de usuario, definidos en el plan del tesauro. Las fuentes de términos pueden ser de dos tipos según Soergel, (2004):

- Fuentes estructuradas: contienen términos seleccionados y estructurados con respecto a algún principio:
    - Listas de descriptores, encabezamientos, clasificaciones u otros tesauros;
    - Nomenclaturas de una disciplina, especialmente si están aprobadas por algún organismo internacional;
    - Tratados terminológicos de una disciplina;
    - Obras lexicográficas: enciclopedias, léxicos, diccionarios, glosarios, etc;
    - Tablas de contenidos e índices de actas de conferencias, libros de texto, manuales y programas o fichas docentes de cursos;
    - Índices de revistas, resúmenes y otras publicaciones y bases de datos;
    - El resultado obtenido por programas de clasificación automática; y
    - Los términos consensuados por un comité de expertos.
- Fuentes abiertas, no limitadas: los términos no están ordenados o tienen que ser inferidos o derivados del contenido como:
    - consultas, peticiones o perfiles de interés de usuarios obtenidos de los registros de históricos de los usuarios de Sistemas de Información o de estudios sobre los intereses de los usuarios, encuestas, entrevistas, etc;
    - memorias de proyectos I+D;
    - índices de colecciones de documentos muestra construidos de forma libre por expertos en el dominio;
    - títulos, resúmenes, textos completos, revisiones de libros, artículos de revistas, comunicaciones, páginas web, documentos de trabajo, etc; y
    - los resultados de búsquedas en la Web.

3) Recolección de términos. Consiste en seleccionar, de las fuentes, los términos con mayor contenido semántico. La recolección se ha hecho, hasta la llegada de los ordenadores personales, de forma completamente manual, leyendo y escogiendo los términos que el especialista consideraba más adecuados. Sin embargo, actualmente, se realiza de modo automático, mediante herramientas de extracción automática de términos (Feldman y Sanger, 2007, Velardi, et. al, 2008, Wikipedia, 2009: http://en.wikipedia.org/wiki/Terminoinology_extraction), o de forma semiautomática, utilizando herramientas de lexicografía computacional (Byrd, et. al., 1987) y

herramientas de análisis textual[4] que preseleccionan, de las fuentes en formato digital[5], los términos susceptibles de incluirse en el tesauro. Sobre estas listas términos se realiza una segunda selección manual más precisa.

4) Análisis y diseño del esquema conceptual. El esquema conceptual define las categorías conceptuales y los tipos de relaciones con las que se va a representar el dominio de conocimiento del tesauro. En los métodos de construcción deductivos, el diseño es la fase que sigue al análisis del dominio que, a su vez, es anterior a la selección de fuentes y a la recolección de términos (figura 7.1. derecha). Un comité de expertos realiza el análisis conceptual del dominio para definir los temas principales en el que se subdivide la disciplina con los que se formarán las categorías principales. Esta operación es una continuación del análisis efectuado en los preliminares, en los que se definió el ámbito y cobertura del tesauro. Un tipo habitual de análisis conceptual es el *análisis por facetas*. Consiste en: a) crear una lista de conceptos básicos, b) organizar los conceptos elementales en facetas, c) organizar cada faceta en jerarquías, d) incluir los conceptos compuestos (intersecciones o especializados) en el marco de las jerarquías. El resultado es un esquema conceptual para organizar los términos formado por: facetas, subfacetas (opcional), términos raíz de jerarquías, tipos de relaciones y microestructura, en caso de considerarse necesaria (figura 7.2).

Figura 7.2. Facetas y términos raíz resultado del análisis por facetas en el tesauro Eurovoc[6].

[4] Las herramientas de análisis textual procesan textos para elaborar listas de palabras con sus frecuencias, colocaciones, listas de palabras clave, distribución espacial de las palabras, etc.
[5] Cuando las fuentes no están digitalizadas y son muy relevantes se escanean o se transcriben
[6] http://europa.eu/eurovoc/

Para evaluar el esquema conceptual, normalmente, se utiliza una muestra de términos suficientemente representativa de todas las categorías, mediante una primera selección de fuentes, recolección y clasificación con, aproximadamente, un 10% del total de términos previstos. Si se encuentran términos que no encajan adecuadamente con el esquema conceptual éste se modifica. El proceso de diseño, recolección de nuevos términos muestra y clasificación se repite (ver línea punteada en la figura 7.1 derecha) hasta que el esquema no sufre más modificaciones. Después se continúa el proceso con la implementación y población del sistema informático que almacenará y gestionará los términos.

Cuando se utilizan métodos inductivos, el análisis conceptual suele estar basado en el *método de agrupación semántica*. Consiste en agrupar los términos en listas o gráficamente en mapas conceptuales bidimensionales, colocando los términos más próximos en significado juntos, de forma que se van generando grupos de términos que constituirán facetas o categorías, dependiendo de si son listas disjuntas o no. La cercanía semántica debe definirse aplicando criterios, reglas o principios previamente estipulados. Por ejemplo, pueden aplicarse principios naturales: (i) orden cronológico, de evolución, de secuencia; o (ii) principios de carácter más conceptual: complejidad creciente, de abstracto a particular, de universal a específico, según la importancia para la indexación y consulta, etc. Dentro de cada grupo se pueden crear subgrupos, formando sub-facetas o sub-categorías. Finalmente, dentro de cada grupo se van estableciendo relaciones semánticas, jerárquicas, de equivalencia o de asociatividad (figura 7.3).

Las reglas para definir los tipos de relaciones entre los términos están recogidas en las normas estándares para la construcción de tesauros (capítulo 4). En los métodos automáticos se utilizan técnicas estadísticas. Si se definen relaciones específicas del dominio, deben representarse también en el esquema conceptual.

El análisis y clasificación conceptual es un proceso inductivo e iterativo. Es necesario repetir el proceso de recolectar, analizar y clasificar términos hasta que el esquema sea estable y representativo del ámbito del tesauro.

Figura 7.3. Agrupaciones semánticas de términos específicos del término photographs en el Tesauro Art and Architecture[7]

5) Implementación y población del tesauro. Una vez que está definido el esquema conceptual éste se implementa en algún modelo de datos (capítulo 5) para poder crear el sistema de almacenamiento y organización de los términos, normalmente una base de datos. Este sistema debe permitir la publicación del tesauro en formato papel o electrónico. La implementación consiste en i) traducir el esquema conceptual a un esquema de datos informático con el que se crea el "esqueleto" o esquema de datos del sistema de almacenamiento, y después ii) poblar el esquema de datos insertando los términos seleccionados. Las relaciones semánticas entre los términos han sido definidas en el esquema conceptual de la fase anterior. En el capítulo 5 se han revisado los modelos de datos informáticos utilizados, con mayor frecuencia, para la construcción de

[7] http://www.getty.edu/research/conducting_research/vocabularies/aat/

tesauros. En el capítulo 6 se ha aplicado el modelo de implementación de datos relacional para crear una base de datos relacional adecuada al modelo conceptual de HL desarrollado y propuesto en este trabajo de tesis. Estos capítulos muestran de forma detallada cómo llevar a cabo esta fase. Los modos de acceso han sido tratados en los capítulos 3 y 4.

6) Publicación. Llamamos publicación a la operación de hacer accesible el contenido del tesauro. Para una publicación electrónica en línea es necesario crear una interfaz entre el usuario y el sistema de gestión y almacenamiento de datos. Esta aplicación suele estar integrada en el sistema de gestión de la base de datos y/o en el sistema de mantenimiento del tesauro.

6) Uso y evaluación. La evaluación es una operación que se realiza de forma permanente durante y al final de la construcción del tesauro por ser un proceso incremental. Durante la construcción, la inserción de términos implica comprobar que se mantiene la coherencia del esquema y contenido del tesauro. Al finalizar, pueden evaluarse los siguientes aspectos:

1. La corrección formal, llevada a cabo por expertos del dominio y lexicógrafos o documentalistas. Consiste en comprobar que:
   1.1. El tesauro incluye los preliminares que documentan el plan de la obra;
   1.2. El tesauro incluye varias formas de presentación, al menos dos, alfabética y jerárquica;
   1.3. Se utilizan correctamente las relaciones semánticas, BT/NT, USE y RT;
   1.4. El tesauro se ajusta a los estándares internacionales;
   1.5. Los términos tienen bien descrito su significado con el contexto, las notas de ámbito o los cualificadores;
   1.6. Las relaciones recíprocas son correctas;
   1.7. Están representados todos los tópicos del dominio; y
   1.8. Los términos son suficientemente descriptivos.

   Una técnica utilizada habitualmente para evaluar los puntos 1.7 y 1.8 consiste en extraer las palabras clave de una muestra aleatoria de fuentes de términos pertenecientes al ámbito del tesauro -en Aitchinson et al. (2000) se sugiere entre 500 y 1000- y no utilizadas como fuente de términos en la construcción del tesauro. Se comprueba si estas palabras forman parte del tesauro.
2. La riqueza relacional del tesauro. Se evalúa según una serie de parámetros que dan cuenta de lo útil que puede ser el tesauro para dirigir al usuario hacia los

- *grado de conexión*: mide la proporción entre los descriptores relacionados y el total de descriptores del vocabulario;
- *accesibilidad o tasa de enriquecimiento*: mide el número de referencias recibidas por los términos descriptores del vocabulario. Una accesibilidad de 2,923 implica que en media cada término está referenciado por, al menos, otros 3 términos del vocabulario. Los valores aconsejados están entre 2 y 5. Valores por encima de 5 indican demasiadas conexiones que dificultan y obscurecen la búsqueda. Valores por debajo indican pocas relaciones semánticas, por lo que el tesauro pierde su eficacia como tal convirtiéndose más en un índice o lista de términos;
- *conectividad*: (b-a)/b, siendo a el número de términos del vocabulario que están sin conectar y b es el total de términos del vocabulario. El tesauro será tanto mejor cuanto más cercano sea este valor a la unidad;
- *tasa de equivalencia*: proporción de no descriptores respecto de descriptores. Es recomendable que sea mayor que 1, es decir que por cada descriptor hay más de un término equivalente, lo que aumenta la probabilidad de encontrar el concepto buscado;
- *tasa de reprocidad*: calcula el número de relaciones TG, TE y RT que mantienen las relaciones recíprocas;
- *ambigüedad:* se calcula con la ecuación (b-a)/b, donde a es el número de descriptores que pueden ser ambiguos porque no tienen notas de ámbito, cualificadotes o relaciones jerárquicas y b es el número total de descriptores del vocabulario;
- *flexibilidad*: es la proporción de palabras que forman parte de descriptores o de no descriptores compuestos (multipalabra). Los valores recomendables deben estar por encima de 0,6;
- *nivel de precoordinación*: es la media del número de palabras por descriptor; y
- *tamaño de las categorías*. Es recomendable entre 30 y 40 términos por categoría.

En Gil (1998a) se presenta un estudio experimental de estos parámetros para la evaluación de seis tesauros del español. Actualmente, el Instituto de Investigaciones

en Humanidades y Ciencias Sociales, tiene en marcha un proyecto de investigación (2008-2010) sobre "el desarrollo de una metodología para la evaluación de tesauros en línea y en español", dirigido por Martínez y Tamayo, A y Ristuccia, C.

3. La usabilidad recoge la utilidad del tesauro desde el punto de vista del usuario. Se aplican métodos cuantitativos, por ejemplo, con encuestas, y/o cualitativos, como los etnográficos, que recogen y estudian el comportamiento del usuario respecto del manejo del tesauro;
4. La ganancia para los sistemas RI mide la mejora en eficiencia que proporciona el tesauro respecto de los parámetros de relevancia y completitud. Esta evaluación se realiza cuando el tesauro tiene la función de refinar las consultas del usuario, expandiéndolas o precisándolas.

7) Mantenimiento

"Un lenguaje de indexación se queda obsoleto tan pronto, o si no antes, de que sea publicado, por lo tanto los tesauros "vivos" deben actualizarse regularmente" (Aitichison, 2000:169).

El mantenimiento es uno de los principales problemas de los tesauros, en primer lugar, porque la lengua está en constante cambio y un tesauro que no recoja estos cambios deja de ser efectivo; y en segundo lugar, porque supone un esfuerzo considerable la actualización. En este sentido, los tesauros muy específicos o los que pertenecen a áreas de conocimiento que están en constante cambio, como las tecnologías, son los que requieren actualizaciones más frecuentes. Por el contrario, los tesauros más generales son más estáticos y requieren menos esfuerzo de mantenimiento.

En cualquier caso, el mantenimiento del tesauro debe realizarse de forma metodológica para evitar inconsistencias o imprecisiones. Normalmente es el editor, con un pequeño equipo de especialistas asesores, quienes deciden qué y cuándo se revisa y actualizan estas obras. Para ello, durante el uso del tesauro, en la indexación o en la búsqueda, se van recogiendo términos candidatos. Estos términos se pueden incorporar al tesauro temporalmente para comprobar su efectividad hasta que se decida cambiar su estado a insertado o borrado. Las operaciones para realizar las modificaciones fueron definidas en el capítulo 4 (sección 4.3.3), junto con la metodología de control de inconsistencias.

La construcción y el mantenimiento del tesauro se realizan con el apoyo de herramientas informáticas de edición (Moya y Gil, 2001). Estas herramientas pueden estar integradas en el Sistema de Recuperación de Información que utiliza el tesauro, pero también existen aplicaciones independientes. Una revisión de las principales

funciones de estas aplicaciones puede encontrarse en (Ganzmann, 1990). Una lista actualizada de Software de Gestión de Tesauros puede consultarse en la página de la American Society for Indexing[8].

7.1.2. La construcción automática

El proceso de construcción inductivo o deductivo de tesauros de la figura 7.1, puede ser automatizado. Aunque los resultados son limitados respecto de la calidad[9] y exhaustividad, desde el punto de vista práctico, los tesauros así obtenidos son muy útiles para clasificar e indexar vastos dominios de información, como por ejemplo, los repositorios de recursos digitalizados, incluido Internet, que por su extensión no son abarcables de forma manual (Chen, et al., 2003). También son útiles como una primera aproximación o fase de construcción de tesauros, que puede ser completada manualmente con un coste menor que si se aborda todo el proceso de forma manual[10] (Aitchison et al., 2000).

Los métodos de construcción automática de tesauros forman parte del área de RI y de la Lingüística de Corpus, y están muy relacionados con las operaciones de indexación automática, generación de resúmenes y clasificación automática (Crouch, 88). Normalmente, se basan en el método inductivo, en el que se aplican técnicas estadísticas y de procesamiento del lenguaje natural (Crouch y Yang, 1992; Grefenstette, 1994; Calzolari, 1994; Yang y Powers, 2008) para extraer, a partir de las fuentes seleccionadas, los términos más representativos del contenido, aplicando técnicas de indexación automática, y las co-apariciones de términos, es decir, grupos de dos o más términos que aparecen juntos con una frecuencia alta, aplicando técnicas de agrupamiento o clustering y técnicas de clasificación automática. En este trabajo no consideramos los métodos estadísticos de construcción de vocabularios, que sirven para obtener los esquemas conceptuales del dominio que organizan los términos del tesauro, porque aplicamos un modelo simbólico general para representación del conocimiento contenido en el tesauro que es el modelo HL presentado en el capítulo 6. La aproximación que se propone se basa en construir automáticamente o manualmente los tesauros aplicando un modelo de contenido, el modelo HL a las estructuras terminológicas existentes en las fuentes de términos. Estas estructuras contienen,

[8] http://www.asindexing.org/site/thessoft.shtml
[9] Entendemos por calidad que el tesauro se ajuste a los criterios de evaluación anteriormente expuestos.
[10] Ver, por ejemplo, la operación de recolección de términos.

explícitamente marcados, los términos, las clasificaciones y/o las relaciones semánticas, por lo que en vez de aplicar técnicas estadísticas o del procesamiento del lenguaje para obtenerlas, se pide al usuario que ha construido las estructuras-t que defina el tipo de estructura-t.

7.2. Una nueva metodología para la construcción de tesauros académicos de explotación

7.2.1. Justificación y premisas

La construcción y uso de tesauros académicos de explotación en formato electrónico surge en la actividad universitaria por la necesidad de disponer de esquemas conceptuales terminológicos sobre una disciplina y entre disciplinas que organicen los conceptos, relacionen conceptos y términos, proporcionen significados precisos para (Soergel, 2002a):

1) ayudar al estudiante a aprender y comprender el lenguaje de especialidad, los términos y conceptos, apoyar el aprendizaje basado en marcos conceptuales, ayudar al estudiante a formular correctamente sus preguntas, ayudar a la lectura y comprensión de los textos, ayudar a la escritura sugiriendo los términos más acertados para expresar las ideas;
2) como herramienta para la creación y explotación de colecciones de recursos didácticos digitalizados: la creación de material didáctico basado en esquemas conceptuales del dominio, de indexación y de búsqueda navegando en las estructuras de clasificación, expandiendo las consultas, orientando la búsqueda hacia conceptos más específicos, genéricos o cercanos semánticamente, e incluso unificando la búsqueda en varias colecciones de recursos; y
3) apoyar al investigador, proporcionando un marco conceptual coherente sobre el que explorar el contexto y dimensiones del problema, definir y estructurar el problema.

Sin embargo, la construcción sistemática de los tesauros conforme a las metodologías recomendadas, deductiva o inductiva, y presentadas en la sección anterior, no es abordable por los equipos docentes por su complejidad, y porque implica una alta inversión económica y/o de tiempo de dedicación.

Además, la experiencia indica que los profesores universitarios no reutilizan los tesauros ya existentes, oficiales o generales, bien por desconocimiento o bien porque no

son útiles para los fines docentes y de investigación[11] (CEN CWA 15453, 2005). Estos tesauros suelen ser demasiado amplios y la disciplina de interés no suele estar completa, o la terminología y estructura conceptual no se corresponde con la que utiliza y necesita el profesor.

La consecuencia es la aparición, en este contexto universitario, de microtesauros más adaptados al lenguaje del especialista y a la estructura de la información y a los recursos que maneja el profesor-investigador. En la práctica estos "tesauros académicos" se construyen ad hoc y de forma libre por los mismos equipos de profesores que los utilizan, e, incluso, participando estudiantes de los últimos años de licenciatura o de doctorado en lo que constituye una actividad más de aprendizaje. El resultado es que estos tesauros, raramente, conforman los estándares, pero siguen un proceso de construcción metodológico. No siguen un esquema conceptual determinado y no están pensados para ampliarse, o integrarse con otros tesauros más específicos o generales. La consecuencia es un producto poco sistemático, difícil de mantener, de reutilizar y de ampliar, que corre el peligro de ser poco útil cuando se quede anticuado o sea insuficiente para un ámbito de aplicación mayor, lo cual es probable que ocurra en poco tiempo.

En este capítulo se presenta una metodología nueva para construir este tipo de tesauros de origen y uso académico. El objetivo es que sirva para facilitar, a los profesores e investigadores, la construcción y gestión sistemática de sus tesauros de especialidad. Para ello se necesitan ciertas premisas que normalmente se dan en el entorno académico universitario:

1. disponer de *contenidos y recursos didácticos y/o de investigación* sobre la disciplina o disciplinas del ámbito del tesauro;
2. que estos materiales hayan sido creados y documentados por los equipos de profesores que necesitan crear el tesauro o, al menos, que se hayan creado *utilizando el mismo lenguaje de especialidad que usan los profesores* y estudiantes) de esa disciplina;
3. que contengan *estructuras de términos (estructuras-t) en semántica libre*. Denominamos estructuras-t a pequeñas redes de términos con relaciones semánticas, una o varias simultáneamente, que no están previamente establecidas, que están inmersas en el contenido o en las descripciones de los

[11] Normalmente son tesauros construidos por documentalistas con fines de gestión documental (Gil, 1998).

materiales educativos, que han sido creadas por uno o varios profesores, especialistas en la disciplina, de forma libre, es decir, por medio de una elección libre[12], que sean reconocibles y que presenten una forma regular, lo que además permite su extracción automática o semiautomática.

Dos ejemplos de estructuras-t se muestran en las figuras 7.4 y 7.5. Estos ejemplos se han tomado de los casos prácticos que han servido para experimentar la nueva metodología.

En la primera figura se muestra una estructura-t que proviene de la ficha de metadatos que documenta un objeto virtual del museo académico CHASQUI. Los términos y sus relaciones han sido elegidos por los profesores que construyen el objeto virtual. La estructura-t, en este caso, consiste en tres categorías llamadas clasificaciones: sección, cultural y zona. La clasificación cultural, a su vez, contiene cuatro categorías, cada una con un término: Área cultural: ANDINA, Subárea cultural: ANDES CENTRALES, Periodo cultural: INTERMINOEDIO ANTIGUO, Cultura: NAZCA.

Figura 7.4. Estructura-t en los metadatos de un objeto virtual del museo académico CHASQUI

La figura 7.5.muestra el contenido de un vocabulario explicativo creado y etiquetado por profesores de la Facultad de Derecho de la UCM, especialistas en propiedad intelectual y cibernética (Flores, et. al., 2009). Se han destacado en

[12] Lo que no implica que sean originales o únicas.

rojo los términos, y las relaciones semánticas que se recogen para construir el tesauro: los términos "obra" y "bases de datos" son términos relacionados, marcados por los entrecomillados y un asterisco, y pertenecen a al campo "jurídico civil"; y

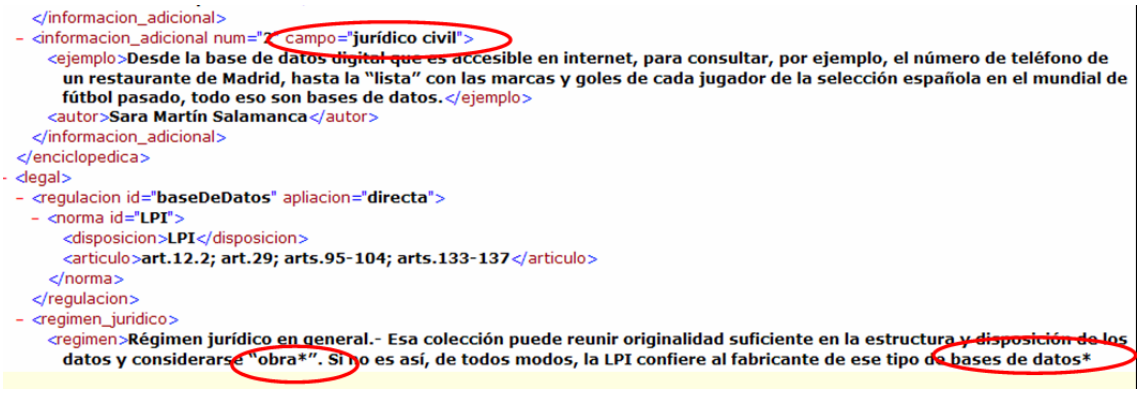


Figura 7.5. Estructuras-t en el contenido (datos) de un vocabulario jurídico

4. finalmente, es necesario *disponer de un modelo general para interpretar y almacenar de forma organizada los componentes de las estructuras-t*. Este modelo es el Higraph Léxico descrito en el capítulo anterior.

A partir de estas estructuras-t, localizadas en los contenidos o en los metadatos de un conjunto de materiales didácticos, se puede diseñar un proceso sistemático de construcción de tesauros.

7.2.2. Descripción del método

El método propone un procedimiento de construcción iterativo e incremental para las etapas previas a la publicación del tesauro (figura 7.6). El proceso comienza con la *identificación y definición* del tipo de estructuras-t contenidas en los recursos didácticos, que son las fuentes de términos. Estas estructuras son una porción del tesauro, una pequeña red de términos y categorías de términos. De forma iterativa se van *extrayendo* las estructuras-t y se van *analizando* conforme el modelo general y único de representación de tesauros de Higraph Léxico (HL). La idea es interpretar los componentes de la estructura-t, que formarán parte del tesauro, respecto de los elementos del HL en categorías, términos, relaciones semánticas, del tipo estándar TG/TE, TR y USE, relaciones de inclusión e, incluso, relaciones específicas del dominio. Si el HL ya está implementado en algún modelo de datos como, por ejemplo, el relacional HL presentado en el capítulo anterior, no es necesaria la fase de implementación, si se utiliza otro modelo de implementación es necesario diseñar, de

nuevo, el esquema de datos. En cualquier caso, consideraremos que ya se dispone de un esquema de datos, meta-esquema informático, para almacenar cada componente de la estructura-t interpretada conforme al modelo HL. Por lo tanto, de forma, progresiva y sistemática, se van *insertando* los componentes de las estructuras-t en una base de datos HL, lo cual va generando el contenido del tesauro, que inicialmente es vacío. El esquema de datos y el contenido del tesauro son dinámicos, puesto que van cambiando conforme se añaden nuevas instancias, que son los componentes de la estructura-t. El tesauro es accesible desde el primer momento, cuando está vacío, ya que siempre pueden hacerse consultas a la base de datos HL. El esquema conceptual del tesauro es la forma gráfica del HL, que puede generarse también realizando las consultas correspondientes en cualquier otro modo de presentación estándar. Este proceso sistemático lo denominamos método HL (método higraph léxico) por ser el modelo HL la base de este proceso. Se compone de siete fases, de las cuales tres requieren de la intervención humana y el resto pueden automatizarse (figura 7.6): 1) identificación y definición del tipo de estructura-t; 2) extracción de estructura-t; 3) análisis e interpretación de las estructuras-t; 4) revisión y adecuación de los términos, categorías y relaciones semánticas; 5) inserción en el HL; 6) publicación del tesauro; y 7) uso y validación.

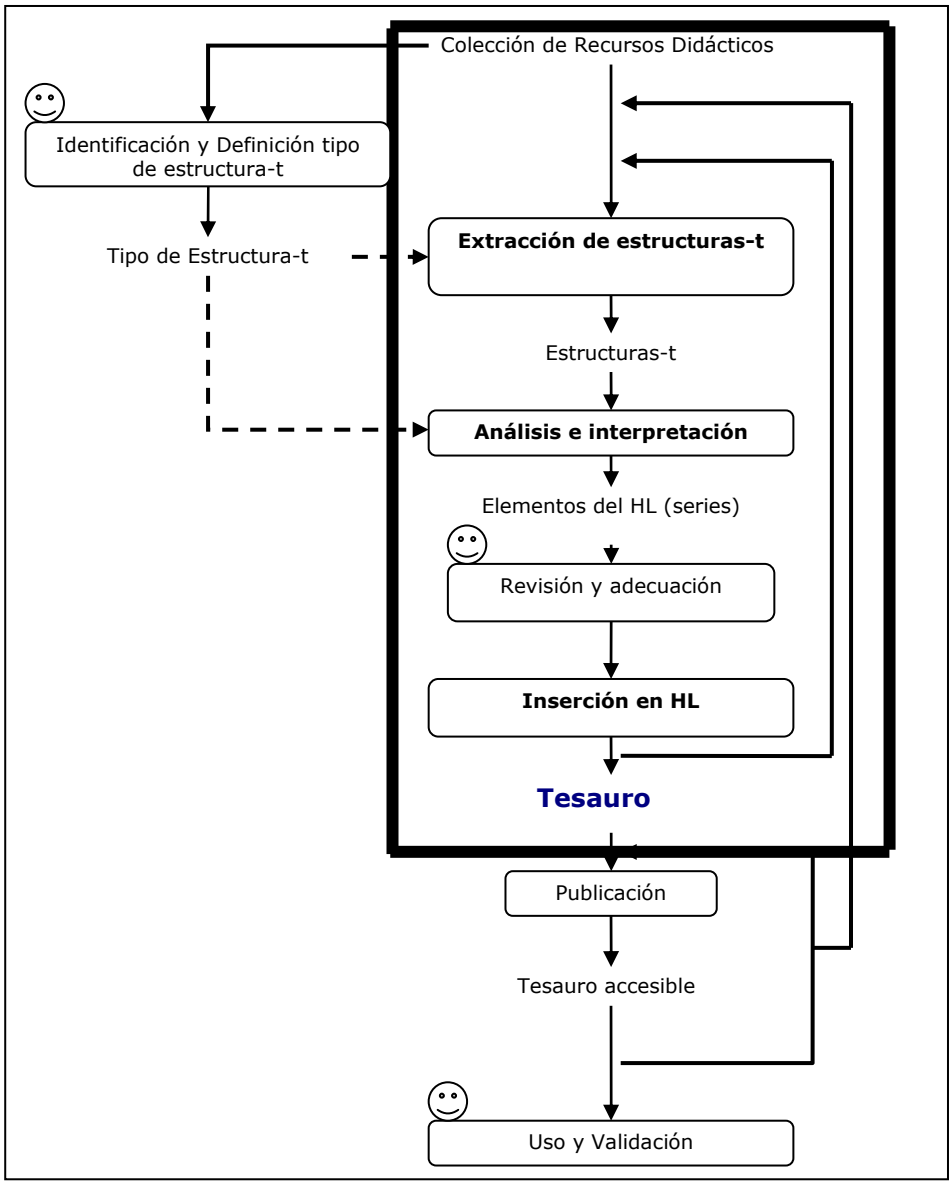


Figura 7.6. Método de HL para la construcción de tesauros[13]

1) Identificación y Definición de las estructuras-t:

Se analizan las fuentes de estructuras-t (metadatos o contenidos didácticos) y se identifican la o las formas de estructuras-t que pueden existir. Esta operación es manual[14] porque los tipos de estructuras-t pueden ser muy variados, incluso utilizando un mismo esquema de metadatos o de contenidos. En esta fase se le pide al usuario que (i) identifique los tipos de redes semánticas de las que consta estructura-t tipo, y (ii) que indique las marcas textuales, normalmente de tipo XML, que distinguen cada estructura-t, cada tipo de red y cada elemento de la red,

[13] Las fases manuales se indican con el símbolo de una cara.
[14] En este trabajo no se ha considerado la posibilidad de automatizarlo aplicando técnicas de PLN o estadísticas.

y su correspondiente elemento HL. La definición y localización del tipo de estructura-t es clave para poder automatizar el resto de las etapas del método HL.

i) Identificación de los tipos de redes y definición del tipo de estructura-t:

una estructura-t es una porción de un HL, por lo tanto está formado por una o varias redes semánticas de tipos diferentes. Las redes semánticas están formados por signos del lenguaje, términos y categorías, y relaciones semánticas, estándares o específicas del dominio. Para definir una red semántica utilizamos un mecanismo de organización similar al que se utiliza en los diccionarios ideológicos o en los sistemas de clasificación automática por agrupamiento: formar series de términos ó categorías asociados por un mismo tipo relación semántica, de forma que, por cada red semántica, habrá una serie de términos y/o categorías. El usuario define el tipo de estructura-t escribiendo los tipos de series que puede contener. Imaginemos un tipo de estructura-t típica formada por a) un grupo de categorías y subcategorías, b) que contienen a su vez términos y c) que estos términos pueden estar relacionados entre sí por relaciones de hiperonimia-holonimia, TG (figura 7.7); el usuario definirá el tipo de estructura-t escribiendo que puede contener tres tipos de series: a) la serie de inclusión de categorías, b) la serie de inclusión de términos en categorías, y c) la serie TG de términos genéricos-específicos (figura 7.7 izquierda). Las instancias de este tipo de estructura-t pueden tener cero una o varias series de cada tipo, es decir, podríamos encontrarnos con una instancia que sólo tuviera una serie del tipo a) de inclusión de categorías (figura 7.7 centro), o bien otra instancia (figura 7.7 derecha) que tuviera una serie, b), de inclusión de términos en una categoría y dos series, c), del tipo TG.

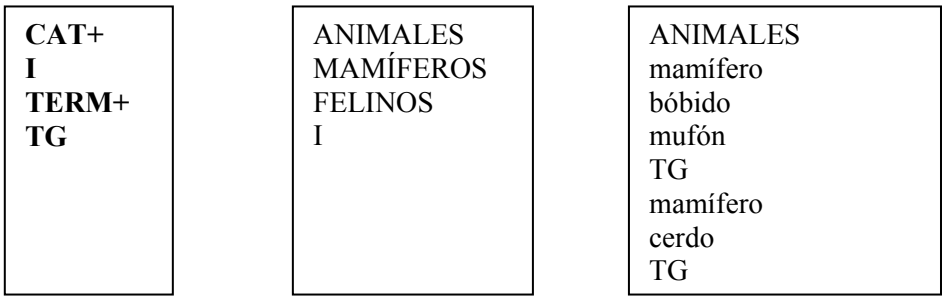


Figura 7.7. Tipo de estructura-t (izquierda) e instancias (centro y derecha)

Para escribir las series hemos creado una nomenclatura, basada en el metalenguaje estándar XML y en el estándar de construcción de tesauros, que es la siguiente[15]:

a) CAT indica categoría;
b) TERM indica término;
c) el símbolo que aparece detrás "*" indica que puede haber 0, 1 o más apariciones del signo que le preceda, el símbolo "+" indica 1 o más, y "?" indica opcional, 0 o 1;
d) I, TG, USE, TR, son los símbolos ya utilizados para indicar las relaciones estándar de inclusión, Término Genérico, Término equivalente preferido y Término relacionado, respectivamente, en cada una de las series; también podrían utilizarse símbolos nuevos introducidos por el usuario para indicar relaciones semánticas específicas del dominio. Utilizamos el convenio de indicar d.1) el tipo de relación de una serie al final de la serie; d.2) escribir en orden los términos, de forma que, en una relación TG, el primero es el término genérico de toda la serie, el segundo es el término genérico del tercer, y así sucesivamente; en una relación USE, el término preferido es el primero; en una relación TR, el primer término se utilizará como base, en general el término inicial será el origen de una relación y el siguiente el destino; y d.3) no es necesario utilizar ningún símbolo para indicar la relación de inclusión entre las series de términos que siguen a una(s) categoría(s); se indica, simplemente, colocando, en primer lugar, la(s) categoría(s), y en segundo lugar, los términos que se incluyen en esa(s) categoría(s); así, por ejemplo, en la figura 7.7 de la derecha, se entiende que todos los términos que aparecen detrás de la categoría ANIMAL están incluidos en ella;
e) se pueden utilizar variables para capturar el texto y para referirnos al valor concreto de esa variable utilizamos, como en XML, &VALOR; y
f) se pueden crear expresiones regulares combinando cualquiera de los metasímbolos anteriores y texto; a estas expresiones las denominamos patrones y sirven para definir la sintaxis de aquellas redes semánticas que aparecen en las estructuras-t escritas de forma regular, por ejemplo, en el caso práctico segundo que presentaremos en el capítulo siguiente, aparece un tipo de red semántica de inclusión de categorías con la forma:

[15] Esta nomenclatura utiliza los símbolos del estándar XML de ocurrencia +, *, ? del estándar XML, y los símbolos estándar de relaciones semánticas de los tesauros.

CAT0/CAT1/CAT2/…/CATn; para definir este tipo de red se puede utilizar el patrón: (CAT0/CAT1)+, que indica que podemos encontrarnos con una serie de dos o más categorías separadas por "/".

ii) Localización de las estructuras-t y sus elementos constituyentes

La segunda cuestión a resolver es indicar dónde y cómo se localizan las estructuras-t y sus componentes. Para ello, en una tabla de correspondencia, el usuario indica qué marcas en el texto o qué expresiones regulares sirven para reconocer una estructura-t o una red semántica (tabla 7.1).

MARCA	**Elemento Estructura-t**
	Estructura-t
	Serie de Categorías
	Relación Categorías
	Serie primera de Términos
	Relación semántica de la serie primera
	Serie segunda de Términos
	Relación semántica de la serie segunda

Tabla 7.1. Tabla de correspondencias organizada en series

En la columna de la derecha se escribe la definición del tipo de estructura-t y en la de la izquierda las marcas o patrones que servirán para reconocerlos. La tabla 7.2 muestra cómo podría ser la tabla de correspondencias del ejemplo de la figura 7.7:

MARCA	Elemento Estructura-t
Marca o patrón	Estructura-t
Marca o patrón de las categorías	CAT+
Marca o patrón de la inclusión de las categorías	I
Marca o patrón del término origen de relación TG	TERM+
Marca o patrón de la relación TG	TG

Tabla 7.2. Organización en la tabla de correspondencias de la estructura-t de la figura 7.7

Este mecanismo de identificación y definición de estructura-t permite la identificación, extracción, análisis e interpretación automática de las estructuras-t. El usuario tiene que construir la tabla de correspondencias, indicando primero las series que forman una estructura-t y luego asociando las marcas o patrones que sirven para reconocer estas series. En los casos prácticos de aplicación del método HL que se presentan en el capítulo 8 se muestran con más detalle tres ejemplos del uso de este mecanismo de identificación y definición de estructuras-t.

2) Extracción de estructuras-t:

A partir de la tabla de correspondencias se pueden reconocer y extraer automáticamente las estructuras-t de las fuentes. El algoritmo de extracción puede resumirse en:

Entrada: fuentes de estructuras-t (txt)
Para cada fuente **hacer**
1) Reconocer la estructura-t utilizando las marcas y patrones de la tabla de correspondencias, y
2) Extraer las estructuras-t
finPara
Salida: conjunto de estructuras-t

El resultado de esta fase será un conjunto de estructuras-t preparadas para ser analizadas e interpretadas en la siguiente fase.

3) Análisis e interpretación de las estructuras-t conforme el modelo HL:

En esta fase se analizan, automáticamente, las estructuras-t definidas y extraídas en las fases anteriores con respecto del modelo HL. El objetivo es reconocer en cada instancia de estructura-t sus componentes HL utilizando el tipo de estructura-t creado en la primera fase. Para ello se aplica una estrategia de análisis en dos etapas sucesivas:

1) obtener las series semánticas de signos que contiene cada estructura-t (figura 7.8), e
2) interpretar cada serie con respecto del modelo semántico HL; en el análisis e interpretación se utiliza la tabla de estructura-t preparada por el usuario para reconocer cada serie y sus elementos: categorías, términos y relaciones semánticas. El procesamiento se realiza modularmente por series siguiendo la secuencia:

Procesar Series de inclusión de CAT, si existen;
Procesar la inclusión de los TERM en CAT, si existen;
Procesar Series de TERM, si existen.

Esta estrategia se resume en el siguiente algoritmo de análisis e interpretación:

Entrada: estructuras-t extraídas de las fuentes
Obtener las series semánticas;
Procesar Series de inclusión de CAT, si existen;
Procesar la inclusión de los TERM en CAT, si existen;
Procesar Series de TERM, si existen;
Salida: tuplas HL

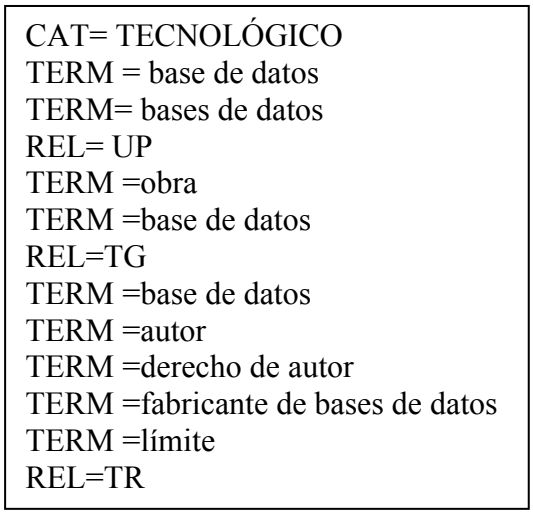


Figura 7.8. Series de signos relacionados

El resultado de procesar una 'Serie' de signos es una interpretación según el modelo HL de la red de signos que forma la serie, es decir, un conjunto de ternas que define la posición diferencial de cada signo en la red (figura 7.9).

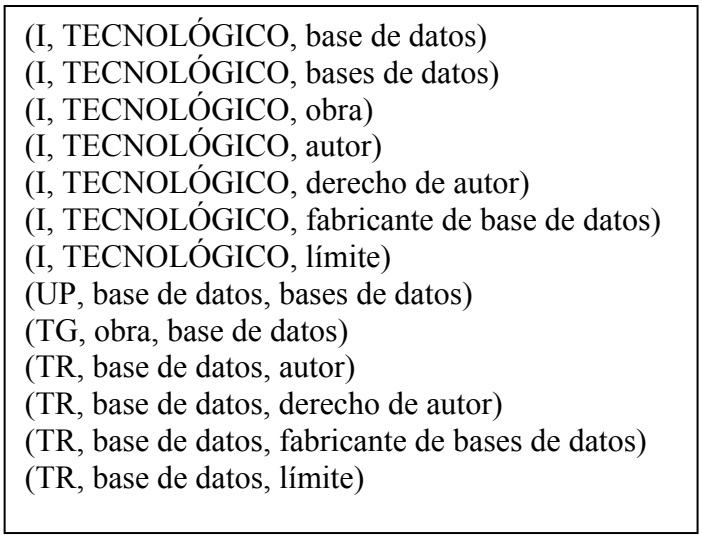


Figura 7.9. Resultado, a nivel conceptual, de la interpretación de la figura 7.8

El algoritmo de procesamiento de una serie, que veremos con más detalle en los casos prácticos, utiliza como gramática el metalenguaje con el que el usuario crea el tipo de estructura-t: básicamente recorre la serie creando una tupla por cada dos signos, con la relación semántica que figura al final de la serie y colocando el primer signo como origen de la relación, es decir, como segundo componente de la tupla y el segundo signo como destino, es decir como tercer componente de la tupla.

Este algoritmo tiene dos versiones dependiendo del nivel de abstracción que se desea en los resultados: a) obtener como resultado las tuplas HL a nivel conceptual (figura 7.9) y dejar que en la siguiente fase, inserción en el HL, el algoritmo de inserción transforme las tuplas conceptuales a estructuras específicas del esquema de implementación de datos del HL antes de insertarlas en el HL del tesauro; u b) obtener, directamente, como resultado las tuplas HL a nivel de implementación de datos, lo que implica que, en la fase siguiente, inserción en el HL, el algoritmo de inserción sólo tiene que ejecutar el resultado de esta fase de análisis e interpretación. La primera opción es interesante en el caso en que se necesite mantener el resultado del análisis de las estructuras-t independiente del modelo de implementación de datos; en el resto de los casos es más efectivo interpretar las estructuras-t directamente con respecto del modelo de implementación de datos. Estas dos opciones las trataremos con detalle en los casos prácticos del próximo capítulo.

4) Revisión y acomodación:

El procedimiento de revisión y acomodación es opcional y manual. Tiene como objetivo corregir errores en las interpretaciones HL obtenidas en la fase anterior y normalizar las formas ortográficas de los términos en el caso en que se desee mantener el control del vocabulario: completar el tesauro con los cualificadores de los términos, las notas de ámbito necesarias y regularizar la forma ortográfica de los términos y categorías. Cuando el tesauro es de tipo postcoordinado –los términos se asignan a los objetos que se van a indexar como claves de recuperación, independientemente de sus relaciones gramaticales- puede prescindirse de esta etapa porque es responsabilidad del software de búsqueda combinar los términos para no perder cobertura[16].

5) Inserción de los datos en el HL:

En esta fase se insertan las tuplas HL obtenidas en la fase tercera, pero teniendo en cuenta:

5.1) la frecuencia de uso para el tratamiento de inconsistencia

El problema de construir inductivamente un tesauro a partir de estructuras-t ad hoc es el control de las inconsistencias. Para resolverlas se propone registrar en la base de datos HL la frecuencia de aparición, contando el número de veces, de los términos, las categorías y las relaciones en las fuentes de términos del tesauro.

[16] Por ejemplo, aplicando reglas de cercanía ortográfica.

Recoger la frecuencia de aparición de los términos y sus relaciones en los tesauros es habitual, porque permite tomar decisiones de mantenimiento sobre la pertinencia de mantener o borrar los términos o los tipos de relación. Cuanto mayor sea la frecuencia de aparición de un elemento del tesauro mayor es su potencial de indexación, clasificación y recuperación.

Este método resuelve las inconsistencias basándose en la suposición de que las estructuras terminológicas que más se repiten son las más útiles y, por lo tanto, son las preferidas. Sin embargo, también suponemos que las estructuras menos frecuentes no deben desaparecer, a no ser que los usuarios las borren explícitamente. Están justificadas porque proceden de una descripción hecha por algún experto, por lo que se considera alternativa, aunque pueda ser errónea. Este tratamiento se puede interpretar de la forma siguiente: "ésta es la estructura preferida por la mayoría hasta este momento, pero hay otros posibles candidatos o alternativas que pueden llegar a ser preferidos según vaya evolucionando el tesauro". Los errores e inconsistencias se consideran, por lo tanto, estructuras alternativas que deben ser valoradas por los expertos de forma manual y explícita para su posible eliminación puesto que, además, implicarán una rectificación en las fuentes de las que proceden las estructuras-t supuestamente erróneas.

De forma más detallada, el tratamiento de inconsistencia según el tipo es:

i) *Caso de inconsistencia en el tipo de relación.* Cuando el tipo de relación entre dos términos no es siempre el mismo: se mantienen ambos y sólo se considerará el tipo más frecuente. Por ejemplo, supongamos que en un HL existe una relación (desarrollo afectivo TR afectividad) y que una nueva estructura-t contiene la relación (desarrollo afectivo TG afectividad), ¿es TR, TG o ambas relaciones la opción correcta? La respuesta es que depende de cuál sea la más utilizada, pero en cualquier caso ambas deben existir, puesto que reproduce una ambigüedad conceptual en la representación del significado de ambos términos que los expertos deben resolver si lo consideran necesario;

ii) *Inconsistencias en las jerarquías TG/TE.* Surgen cuando un camino jerárquico debe insertarse en una jerarquía del HL ya existente, y además ocurre alguna de las siguientes circunstancias:

    a) Al menos un término genérico y específico están invertidos respecto de la jerarquía: la nueva inserción tienen el sentido de la relación TG invertido respecto a la existente en el HL (figura 7.10). Por ejemplo,

supongamos que el término Acción moral (en amarillo) es padre de Ética (en azul), y en la nueva inserción aparece al revés: Ética padre de Acción moral. La nueva relación incompatible se inserta, pero se ajustan las frecuencias de uso. La relación padre-hijo que tenga menor frecuencia se considera incompatible (ver figura 7.10 arco en rojo). Si tienen la misma frecuencia se consideran las dos hasta que una nueva inserción confirme a una de ellas[17].

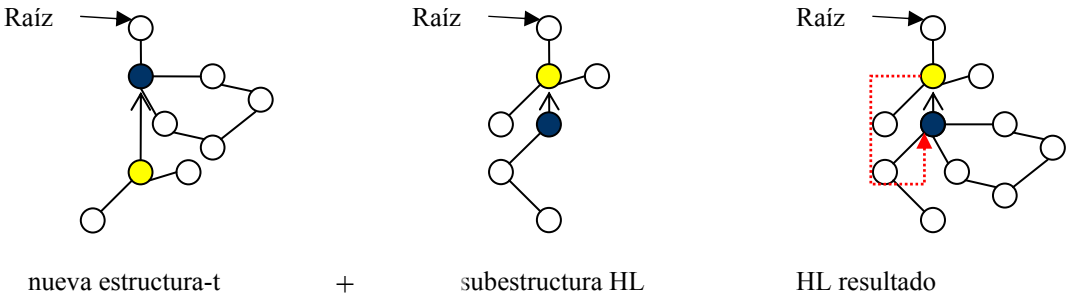


Figura 7.10. Resolución de una inconsistencia padre-hijo

b) Una nueva raíz. Cuando dos términos padre-hijo son raíces de una misma taxonomía (figura 7.11). En este caso, al menos un experto ha considerado que el nodo hijo –en azul- es el más genérico en la categoría-taxonomía, mientras que otro u otros expertos consideran que existe otro término –en amarillo- con significado más amplio. De nuevo la solución es considerar ambas posibilidades, marcando como inconsistente la que menor frecuencia tenga.

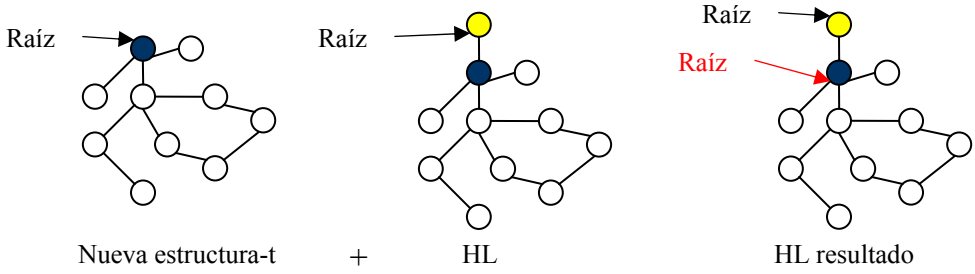


Figura 7.11. Inconsistencia de raíz en una misma categoría

c) Nuevos términos intermedios. Si los dos caminos que hay que combinar en una misma jerarquía tienen longitudes diferentes, en el camino de mayor

[17] En el modelo relacional esto significa que en la tabla hl_macro aparecen dos tuplas TG con los nodos intercambiados y la tupla inconsistente tiene un valor de frecuencia menor.

longitud existen términos nuevos entre algunos pares de términos padre-hijo (figura 7.12)[18]. La inclusión de los nuevos pares no conduce realmente a una inconsistencia puesto que la relación TG es transitiva. Pero explicitar la transitividad es redundante y hace más difícil procesar la jerarquía. Por lo tanto, para evitar la redundancia, se borra la relación directa (A,B), y se insertan dos pares nuevos (A,C), (C,B).

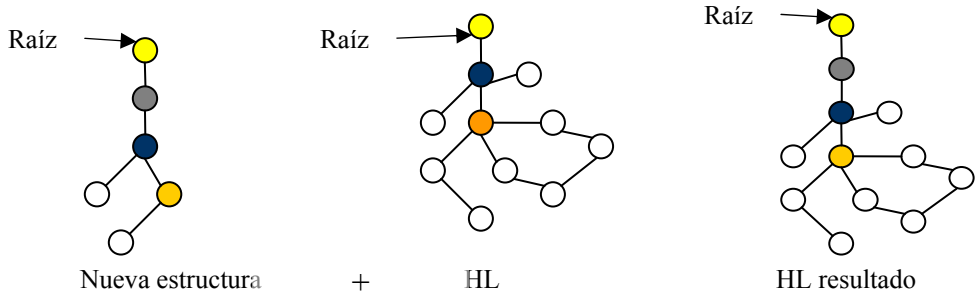


Figura 7.12. Nuevos términos intermedios

5.2) El procedimiento de inserción, teniendo en cuenta el mecanismo de tratamiento de inconsistencias, consiste en añadir al HL las tuplas obtenidas del análisis e interpretación de la fase tercera. En caso de que surja incompatibilidad entre dos tuplas, se comparan las frecuencias y considera como "inconsistente" o "menos deseable", la que menor frecuencia tenga, y si son iguales, podría utilizarse cualquier criterio, por ejemplo no considerar a la más reciente[19]. Sólo hay que tener cuidado con los nuevos términos intermedios, que deben insertarse después de borrar la tupla con los términos iniciales.

Cuando el HL se implementa con el esquema relacional HL[20], las inserciones se expresan en el lenguaje SQL. Básicamente se realizan dos tipos de inserción, una en la tabla hl_micro, y la otra en la tabla hl_macro, de la forma siguiente:

- Tabla hl:micro: la actualización de la lista de términos y categorías. Esta tabla debe tener una columna más para guardar la frecuencia. Todos los términos nuevos se insertan en la tabla. El SGBDs asigna automáticamente un identificador único a cada término y establece su frecuencia de uso en 1, que es el valor inicial por defecto. Si los términos ya estaban en la tabla,

[18] Por ejemplo, en la página 5 del tesauro ETB-LRE en la jerarquía con raíz 'aprendizaje' aparece un nuevo término, 'proceso de aprendizaje', entre la raíz y 'aprendizaje por experiencia'.
[19] Se da preferencia a la antigüedad de una estructura frente a otra porque se valora el tiempo que lleva disponible en el tesauro y que posiblemente ha sido visualizada en los procesos de búsqueda, selección y exploración.
[20] El esquema relacional de datos presentado en la sección de implementación relacional del capítulo 6.

se incrementa en uno su frecuencia de uso. Puede ocurrir que un término nuevo sea una variante ortográfica de otro, pero en este caso también se inserta, incluso aunque sea un error. En las fases de revisión y validación, si el usuario lo considera conveniente, puede borrarlo o mantenerlo como un término equivalente, relacionándolo con un descriptor, con lo cual implica la creación en hl_macro de una nueva relación USE.

- Tabla hl_macro. La actualización de la lista de relaciones. La tabla hl_macro también tiene una columna más que almacena la frecuencia de las relaciones. Se añaden todas las tuplas que representan relaciones entre los signos (TipoRelación, Signo1, Signo2) y se incrementa la frecuencia o la establece a uno si es nueva. Debe tenerse en cuenta los mecanismos de resolución de los distintos tipos de incompatibilidades definidos anteriormente.

Las etapas de extracción estructura-t, análisis e interpretación, e inserción se repiten para cada estructura-t y van "poblando" el HL.

La publicación, uso y validación no forman parte del método en sí, aunque son parte del proceso global de construcción del tesauro. En estas fases hay que tener en cuenta algunas cuestiones relativas al método.

6) Publicación:

El tesauro publicado es el resultado de interpretar el HL, tal y como se definió en el capítulo 6 (sección 3), pero añadiendo la interpretación de las frecuencias: sólo se consideran, en caso de inconsistencia, las tuplas más frecuentes. Las tuplas inconsistentes, como hemos visto, no se borran y permanecen "invisibles" hasta que se validen con nuevas entradas o se eliminen en la etapa de revisión.

7) Uso y validación:

El procedimiento de validación se aplica de forma manual y continua. Si durante el uso del tesauro se detecta un error, omisión e inconsistencias se corrige en el tesauro y, también, en el lugar de procedencia de las estructuras-t, los metadatos o materiales educativos. De esta forma cuanto más se usen los tesauros más consistentes y completos serán estos tesauros y las colecciones de recursos y materiales didácticos que tengan asociados.

7.3. Resumen y conclusiones del capítulo

El método HL aporta una solución específicamente orientada al contexto académico para la construcción y actualización permanente de los tesauros de especialidad, reutilizando las estructuras-t conceptuales que contienen los materiales didácticos creados por los equipos docentes. Se apoya en el modelo general HL para insertar, de forma incremental, las estructuras-t en un esquema de HL básico. Este esquema de datos básico se ha ampliado añadiendo a cada elemento del HL un número que indica el número de veces que este elemento se ha insertado en el esquema básico, es decir, la frecuencia con la que se utiliza en las estructuras-t. La frecuencia sirve para resolver inconsistencias dando preferencia a los elementos con mayor frecuencia respecto de los de menor frecuencia en caso de incompatibilidad.

La ventaja de este método con respecto de los métodos tradicionales deductivos o inductivos es que no es necesario realizar un análisis del dominio, más allá de seleccionar los contenidos o recursos que se quieren explotar con el tesauro, ni diseñar un esquema conceptual del ámbito del tesauros, puesto que éste surge inductivamente, en forma de HL, conforme va poblándose el esquema de datos básico con las estructuras-t. Además, el proceso construye el tesauro de forma automática, excepto en las etapas iniciales y en las dos etapas de revisión y validación, que necesita de la intervención del usuario. En consecuencia: (1) este método sirve de base para la posible construcción de una herramienta software que facilite la construcción de los tesauros académicos de explotación; (2) con la aplicación de este método el tesauro se va creando de forma continua, en un proceso que sólo termina si se acaban las fuentes de términos que son las estructuras-t, este enfoque metodológico se adapta mejor a la naturaleza cambiante de los tesauros que los enfoques tradicionales que tienen una etapa inicial y final determinadas; (3) como el tesauro tiene, en cada momento, un esquema conceptual y un contenido ajustado a la colección de estructuras-t que son fuente del tesauro, este enfoque metodológico garantiza la definición precisa en el lenguaje de los autores del dominio de información o de recursos digitalizados que son fuente de las estructuras-t. Finalmente, (4) como el tesauro se va creando sin intervención de los autores, salvo para definir al comienzo el tipo de estructura-t y realizar las correcciones, este enfoque metodológico define el procedimiento de funcionamiento de una herramienta software que, aplicada a cualquier colección de recursos o materiales

didácticos digitalizados que contengan estructuras-t, sirva para que los profesores e investigadores puedan construir automáticamente los tesauros académicos de explotación de sus colecciones de recursos y materiales académicos.

Capítulo 8

Casos prácticos

En este capítulo se presentan tres ejemplos de aplicación del modelo y método HL para la construcción de tesauros académicos de especialidad. Aunque se basan en desarrollos reales que no se hicieron teniendo en cuenta las ideas propuestas en este trabajo de tesis, pero que han dado como fruto esta investigación, nos han servido para verificar la viabilidad del modelo y método HL y para construir, a posteriori, un prototipo de tesauro. Este proceso de ingeniería inversa[1] también ha servido para estudiar las ventajas e inconvenientes de este método respecto de los métodos generales. En los dos primeros casos tratados se aplica el método para (i) especializar y (ii) crear tesauros académicos de explotación de colecciones de recursos educativos; mientras que en el tercer caso se construye (iii) un tesauro académico de explotación de contenidos especializados con fines didácticos y de investigación.

8.1. La especialización de tesauros generales

8.1.1. Introducción

Este caso tiene como objetivo mostrar una de las aplicaciones académicas en las que el método HL sería de utilidad docente e investigadora: la construcción de tesauros de especialidad a partir de tesauros de referencia generales. La ventaja de especializar los tesauros generales de referencia está en que se mantiene el marco de indexación y clasificación general que proporciona el tesauro de referencia, permitiendo la integración de las colecciones de recursos, más específicas, en sistemas de búsqueda de recursos generales. Además, la especialización, gracias a la ampliación del tesauro general con el método de estructuras-t, ayuda a resolver el problema de imprecisión y de falta de completitud de los tesauros generales cuando se aplican a dominios específicos, como las colecciones de recursos didácticos de una disciplina. El método HL va a permitir añadir exactamente aquellas estructuras semánticas de términos que los profesores necesitan para clasificar sus materiales didácticos y que no están en el tesauro de referencia.

[1] Aplicar ingeniería inversa supone profundizar en el estudio de su funcionamiento, hasta el punto de que podemos llegar a entender, modificar, y mejorar dicho modo de funcionamiento (http://es.wikipedia.org/wiki/Ingenier%C3%ADa_inversa).

Este tipo de aplicación puede llevar a cabo siempre que se cumplan los siguientes supuestos:

- se dispone de un tesauro general de referencia, para clasificar una determinada colección de recursos educativos, pero es insuficiente o no está bien ajustado a la colección;
- se va a utilizar el tesauro como un mapa conceptual para indexar y clasificar los recursos didácticos y para guiar en la búsqueda y selección de dichos recursos;
- los profesores, y eventualmente los estudiantes, han clasificado los recursos o materiales didácticos con los términos del tesauro general de referencia siempre que sea posible y, solamente en caso de que no sean adecuados los términos o las relaciones semánticas de términos para clasificar los recursos, se introduce su propia clasificación; y
- el esquema de implementación de datos del tesauro de referencia y del HL debe ser el mismo; en este caso en particular, suponemos que el tesauro de referencia está implementado en la base de datos relacional propuesta en el capítulo 6.

Este caso práctico se basa en (i) el uso del tesauro general ETB en español para clasificar los recursos didácticos digitalizados en los repositorios europeos y en el, recientemente creado, repositorio español AGREGA; (ii) en la especialización realizada, manualmente, en el tesauro ETB por un comité de expertos de AENOR para adaptar el tesauro a las materias de enseñanza en España; y (iii) en el uso de la versión española del modelo de metadatos IEEE-LOM, recientemente definida por AENOR: LOM-ES v1.0 (LOM-ES, 2008) para escribir las estructuras-t.

8.1.2. Utilización del tesauro de referencia ETB en español

El tesauro ETB es uno de los vocabularios recomendados y utilizados para la clasificación de recursos educativos con el modelo de metadatos IEEE-LOM. Se creó en el contexto del proyecto European Treasure Browser, para facilitar el intercambio de información entre los distintos repositorios de información sobre educación en Europa Es un tesauro multilingüe disponible en 14 idiomas: alemán, albanés, árabe, danés, español, finlandés, francés, griego, hebreo, holandés, húngaro, inglés, italiano y sueco. El contenido del tesauro se presenta en un documento pdf con tres índices: uno alfabético, que contiene la información sobre las relaciones semánticas de cada término y la traducción del mismo a las otras 13 lenguas en las que está disponible el tesauro; otro permutado KWIC, en el que se incluyen los reenvíos a los términos admitidos en el

caso de los no descriptores; y otro jerárquico, en el que se incluyen también los términos relacionados. El número de descriptores es de 1.155 y el de no descriptores varía en función de la lengua (Trigari, 2002; Monchon y Sorli, 2007). El tesauro ETB aporta un marco común para la indexación y búsqueda de recursos educativos en Europa independiente de la lengua.

La versión española está disponible en línea desde el año 2006, http://www.r020.com.ar/etb/index.php. Esta versión tiene una extensión de 1477 términos, de los que 300 son no descriptores, y 1546 relaciones semánticas. El tesauro está formado por 17 categorías principales y dos de estas categorías están divididas en subcategorías. Cada categoría contiene un conjunto de términos organizados en jerarquías de especialización/generalización. Además contiene redes de términos relacionados (TR), que pueden involucrar términos situados en categorías diferentes (figura 8.1).

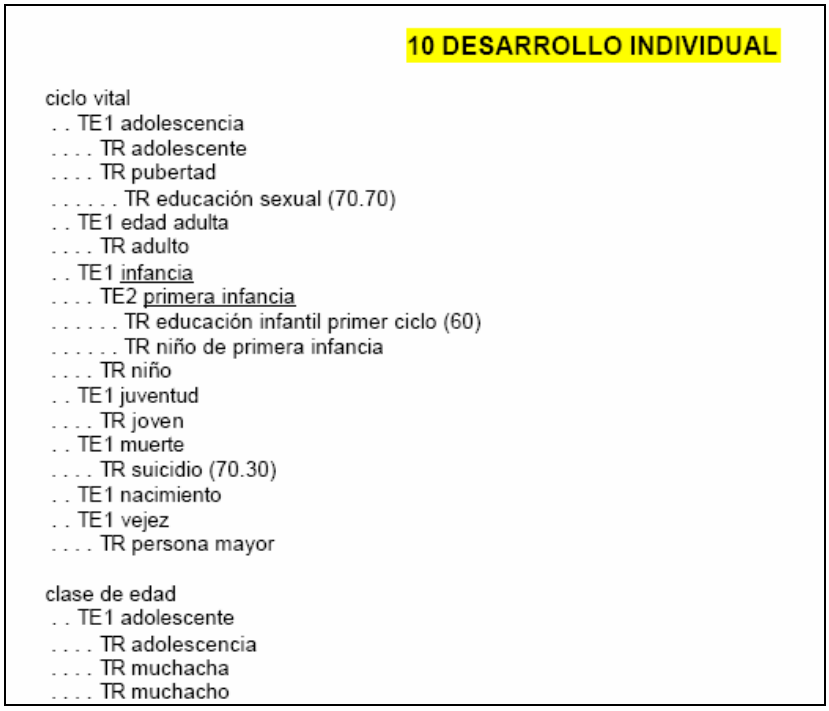


Figura 8.1. Estructura sistemática del tesauro español ETB

Esta versión española ha sido recientemente revisada en el marco del proyecto AGREGA, para adaptarla a la clasificación de recursos educativos de educación primaria y secundaria en España[2]. El objetivo del proyecto AGREGA es crear, en España, una red de repositorios de recursos educativos. Estos recursos se documentan utilizando un nuevo perfil del estándar IEEE-LOM de metadatos educativos llamado

[2] http://www.proyectoagrega.es/

LOM-ES (LOM-ES, 2008). Este perfil recomienda utilizar los términos de la nueva versión del tesauro ETB, llamada ETB-LRE MEC-CCAA v.1.0, para abreviar tesauro ETB-LRE (Berrocal et al., 2008).

Disponer de estas dos versiones ETB (2006) y ETB-LRE (2008) da la posibilidad de poder evaluar la efectividad del método inductivo HL para especializar el tesauro general ETB (2006) al dominio específico de los objetos virtuales del repositorio AGREGA. Este caso práctico reproduce las modificaciones realizadas manualmente por el comité de expertos del Grupo de Trabajo 9 perteneciente al Subcomité 36 de Tecnologías de la Información y la Comunicación para el aprendizaje (SC36) de la Asociación Española de Normalización y Certificación (AENOR). Con estas modificaciones hemos creado las clasificaciones LOM-ES que son la fuente de estructuras-t para especializar el tesauro ETB (2006). Las modificaciones están marcadas con un código de colores, en el propio tesauro ETB-LRE, por lo que ha sido sencillo localizarlas, tipificarlas y utilizarlas como nuevas estructuras-t (figura 8.2). Son de dos tipos:

1.- taxones nuevos que especializan el contenido (TE1 o superior y TR)[3];

2.- taxones eliminados, sustituidos o modificados en su nivel de profundidad (TE1 o superior y TR) en color rojo.

El método HL, siguiendo la estrategia de conservar todas las estructuras-t, no elimina estos taxones, por lo que las modificaciones son, siempre, del tipo insertar nuevos taxones. Se tomará como ejemplo ilustrativo el primer tipo, taxones nuevos que especializa el contenido. En concreto, se mostrará cómo realizar la inserción de los términos *Acción Moral*, *Fundamento antropológico* y *Ámbito Moral* del tesauro ETB-LRE que no están en el tesauro ETB (2006).

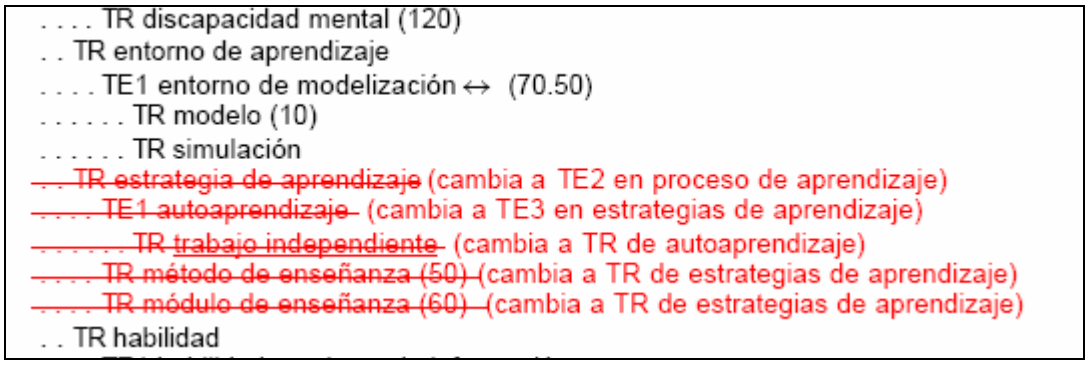


Figura 8.2. Modificaciones en la versión española del tesauro ETB

[3] TE1 se refiere a Término Específico de nivel 1.

8.1.3. Aplicación del método

El procedimiento comienza con el tesauro de referencia ETB (2006) implementado como un HL y un conjunto de clasificaciones LOM-ES que contienen los términos, relaciones semánticas y categorías nuevas del ETB-LRE (2008). Estas clasificaciones son las modificaciones que han introducido manualmente el comité de expertos de AENOR.

Primera fase: identificación y definición del tipo de estructuras-t

Las estructuras-t son caminos taxonómicos[4] (<taxonPath>), que están incluidos en las clasificaciones, escritas con XML, de los metadatos LOM-ES de los recursos didácticos (figura 8.3). Para crear un tesauro académico de explotación, sólo son de interés las clasificaciones que describen conceptualmente un recurso, que son aquellas cuyo propósito –marcado con la etiqueta *purpose,* según se recomienda en LOM-ES- es *disciplina*, o *idea*. Estas clasificaciones son pequeñas redes de términos fácilmente identificables y presentan una forma regular que permite su interpretación y extracción automática.

El tipo de estructura-t es siempre un término dentro de una categoría o una jerarquía TG/TE de términos dentro de una categoría (figura 8.3). Como LOM-ES sólo permite caminos taxonómicos, no aparecen redes de términos asociativas o equivalentes.

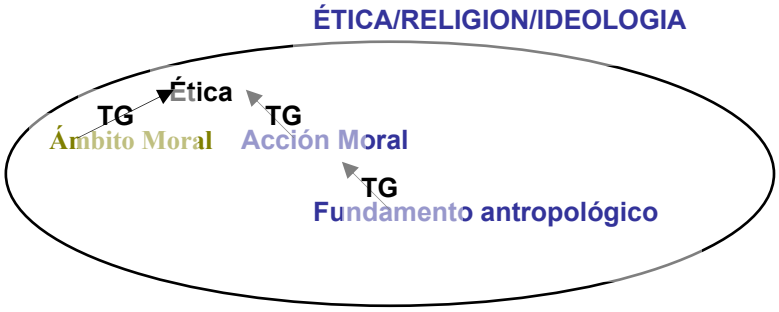


Figura 8.3. Representación gráfica del tipo de estructura-t

La figura 8.4 muestra el código XML donde están insertadas las estructuras-t representadas gráfica en la figura 8.3. Un camino taxonómico incluye varios taxones, el primero es la categoría y los siguientes son los términos incluidos en dicha categoría.

La categoría se identifica con un valor único que aparece entre las etiquetas <id> y </id>: M70.90 y un nombre de categoría que aparece en la etiqueta <entry>, subetiqueta <string>: ÉTICA/RELIGIÓN/IDEOLOGÍA.

[4] Un camino taxonómico es un conjunto de términos organizados del más genérico al más específico.

```
<classification>
<purpose>
        <source uniqueElementName="source">LOM-ESv1.0</source>
        <value uniqueElementName="value">disciplina</value>
</purpose>
<taxonPath>
        <source>
        <string language="es">ETB-LRE MEC-CCAA V.1.0</string>
        48/62 – LOM-ES v.1.0 12/03/2008
        </source>
        <taxon>
        <id>M70.90</id>
        <entry>
        <string language="es">Ética/Religión/Ideología</string>
        </entry>
        </taxon>

        <taxon>
        <id>441</id>
        <entry>
        <string language="es">Ética</string>
        </entry>
        </taxon>

        <taxon>
        <id>3597</id>
        <entry>
        <string language="es">Acción Moral</string>
        </entry>
        </taxon>

        <taxon>
        <id>1576</id>
        <entry>
        <string language="es">Fundamento antropológico</string>
        </entry>
        </taxon>
</taxonPath>
<taxonPath>
        <source>
        <string language="es">ETB-LRE MEC-CCAA V.1.0</string>
        </source>

        <taxon>
        <id>M70.90</id>
        <entry>
        <string language="es">Ética/Religión/Ideología</string>
        </entry>
        </taxon>
        <taxon>
        <id>441</id>
        <entry>
        <string language="es">Ética</string>
        </entry>
        </taxon>
        <taxon>
        <id>3601</id>
        <entry>
        <string language="es">Ámbito Moral</string>
        </entry>
        </taxon>
</taxonPath>
```

```
<description>
<string language="es">Recurso sobre el Ámbito, fundamento y especificidad de la Ética como
disciplina filosófica</string>
</description>
<keyword>
<string language="es">Fundamento de la moral</string>
</keyword>
<keyword>
<string language="es">Especificidad de la moral</string>
</keyword>
<keyword>
<string language="es">Filosofía moral</string>
</keyword>
</classification>
```

Figura 8.4. Taxones dentro de una clasificación de un recurso (fuente lom_esv1)

Los siguientes taxones son términos y sus identificadores únicos, siguiendo el orden de más general a más específico:

```
<taxon>
<id>441</id>
<entry>
<string language="es">Ética</string>
</entry>
</taxon>

<taxon>
<id>3597</id>
<entry>
<string language="es">Acción Moral</string>
</entry>
</taxon>

<taxon>
<id>1576</id>
<entry>
<string language="es">Fundamento antropológico</string>
</entry>
</taxon>
```

Con este conocimiento sobre los metadatos LOM-ES y las estructucturas-t que tiene el profesor, porque supuestamente es el autor de las estructuras-t, puede definir el tipo de estructura-t mediante la tabla de correspondencias 8.1.

MARCA/PATRÓN	**Componente**
Etiqueta <taxonpath> de las clasificaciones cuyo propósito (etiqueta <purpose>) es “disciplina”	Estructura-t
Primera etiqueta <taxon><entry>	CAT
Segunda y siguientes etiquetas <taxon><entry>	TERM+
implícito[5]	TG

Tabla 8.1. Tabla de correspondencias para definir el tipo de estructura-t

[5] Si fuera otra debería marcarse, pero en el modelo LOM-ES no se consideran más relaciones que las jerárquicas de generalización.

Segunda fase: extracción de estructuras-t

Esta fase consiste en extraer las estructuras-t del código XML de los metadatos de los recursos didácticos utilizando el conocimiento de la tabla 8.1 y el algoritmo descrito en el capítulo 7:

Entrada: fuente archivos metadata.xml
Para cada fuente:
1) Reconocer la estructura-t en el archivo de metadatos utilizando las marcas y patrones de la tabla de correspondencias, y
2) Extraer las estructuras-t
Fin bucle Para
Salida: conjunto de caminos taxonómicos (taxonPath)

Tercera fase: análisis e interpretación de las estructuras-t en estructuras del HL

En este caso práctico hemos optado por realizar el análisis a nivel conceptual, que como hemos visto en el capítulo 7 se realiza en dos pasos:

1. Obtener las series semánticas de signos, interpretando cada estructura-t extraída conforme la tabla de correspondencias definida por el usuario, y
2. Interpretar las series semánticas conforme el modelo HL, utilizando el metalenguaje.

Aplicando el paso 1 a las dos estructuras-t de la figura 8.4 se obtendrían las series:

CAT= Ética/Religión/Ideología
TERM0= Ética
TERM1= Acción Moral
TERM2 = Fundamento Antropológico
REL= TG
CAT0= Ética/Religión/Ideología
TERM0= Ética
TERM1= Ámbito Moral
REL= TG

Aplicando el paso 2 a estas series, se obtiene la interpretación HL siguiente:

(I, ÉTICA/RELIGIÓN/IDEOLOGÍA, Ética),
(I, ÉTICA/RELIGIÓN/IDEOLOGÍA, Acción Moral),
(I, ÉTICA/RELIGIÓN/IDEOLOGÍA, Fundamento antropológico),
(TG, Ética, Acción Moral),
(TG, Acción Moral, Fundamento antropológico),
(I, ÉTICA/RELIGIÓN/IDEOLOGÍA, Ámbito Moral)
(TG, Ética, Ámbito Moral),

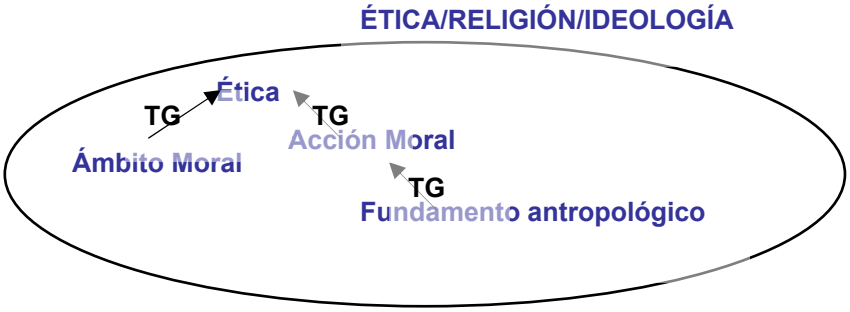


Figura 8.5. Interpretación gráfica del resultado de analizar las estructuras-t de la figura 8.4

Cuarta etapa: revisión y adecuación

En esta fase el usuario puede revisar, de forma manual, las ternas obtenidas en la fase anterior. El objetivo es corregir posibles errores o controlar el lenguaje del tesauro, por ejemplo, aplicando la normalización española (UNE 50106, 1990) actualmente en vigor para la construcción de tesauros monolingües. Normalmente la aplicación de estas reglas se acuerda entre los autores del tesauro. Así, en el tesauro ETB-LRE se aplica la regla de regularización ortográfica siguiente: los términos se escriben con minúsculas y las categorías con mayúsculas, y la regla de regularización morfológica de que todos los signos van en singular. Esta revisión y adecuación puede realizarse también durante la última fase del proceso de construcción, es decir, en la validación.

Cuando estas reglas se aplican a los términos obtenidos en la etapa anterior, el resultado es un archivo con estructuras-t normalizadas que son la entrada para la siguiente fase.

Quinta fase: inserción en el HL

Las interpretaciones obtenidas, que son tuplas del esquema relacional HL, se insertan en la base de datos HL utilizando el lenguaje de gestión de datos relacional SQL:

```
INSERT IGNORE INTO hl_micro (id, signo, tipo_signo) VALUES
(70.90, ÉTICA/RELIGIÓN/IDEOLOGÍA, categoría), (441,'Ética',termino), (3597,'Acción
Moral',termino), (1576,'Fundamento antropológico',termino), (3601,'Ámbito
Moral',termino);
```

La tabla 8.2 muestra el resultado, en la tabla *hl_micro,* de ejecutar esta sentencia SQL. El SGBDs asigna automáticamente un identificador único a cada nuevo término o categoría y establece su frecuencia de uso en 1, que es el valor inicial por defecto. En el ejemplo son términos nuevos *acción moral* y *fundamentos antropológicos*.

Cuando los signos ya están en la tabla, simplemente se incrementa en uno su frecuencia de uso. En el ejemplo, el término *Ética* ya estaba en la tabla y, por lo tanto, la inserción sólo ha cambiado su frecuencia a 2. Puede ocurrir que el término nuevo sea una variante de otro, pero en este caso también se inserta, incluso aunque sea un error. En las

siguientes etapas de revisión y valoración, si el usuario lo considera conveniente, se pueden asignar una forma preferida, con lo cual crea automáticamente una nueva relación USE en la tabla macro.

id	signo	tipo	freq
70	CONTENIDO DE LA EDUCACIÓN	categoría	5
441	Ética	término	2
3597	Acción moral	término	1
1576	Fundamentos antropológicos	término	1
3601	Ámbito moral	Término	2
70.90	ÉTICA/RELIGIÓN/IDEOLOGÍA	término	4

Tabla 8.2. Inserción de signos en la tabla hl_micro

INSERT IGNORE INTO hl_macro (tipo_relación, signo1, signo2)) VALUES
(I, ÉTICA/RELIGIÓN/IDEOLOGÍA, Ética),
(I, ÉTICA/RELIGIÓN/IDEOLOGÍA, Acción moral),
(I, ÉTICA/RELIGIÓN/IDEOLOGÍA, Fundamento antropológico),
(I, ÉTICA/RELIGIÓN/IDEOLOGÍA, Ámbito moral),
(TG, Ética, Acción moral),
(TG, Acción moral, Fundamento antropológico)
(TG, Ética, Ámbito moral);

Las 7 últimas filas de la tabla 8.3 son las nuevas inserciones. La inserción se realiza de la forma prevista: se incluyen las nuevas tuplas, se asigna un identificador a cada una y se asigna 1 a la frecuencia de uso. En caso de que las tuplas ya existiesen, como el caso de la 4ª fila, sólo se incrementa la frecuencia.

Tipo_relación	signo1	signo2	freq
I	CONTENIDO DE LA EDUCACIÓN	HUMANIDADES	1
I	HUMANIDADES	Filosofía	1
TG	Filosofía	Ética	1
I	CONTENIDO DE LA EDUCACIÓN	ÉTICA/RELIGIÓN/IDEOLOGÍA	4
I	ÉTICA/RELIGIÓN/IDEOLOGÍA	Ética	1
I	ÉTICA/RELIGIÓN/IDEOLOGÍA	Acción moral	1
I	ÉTICA/RELIGIÓN/IDEOLOGÍA	Ámbito moral	1
I	ÉTICA/RELIGIÓN/IDEOLOGÍA	Fundamento antropológico	1
TG	Ética	Acción moral	1
TG	Acción Moral	Fundamento antropológico	1
TG	Ética	Acción moral	1

Tabla 8.3. Inserciones en la tabla hl_macro

En este ejemplo no se han detectado inconsistencias, pero resulta interesante destacar que el término *ética*, ya existía en la categoría HUMANIDADES y dentro de la jerarquía con raíz filosofía. La nueva inserción genera una nueva jerarquía con raíz *ética*

en la categoría ÉTICA/RELIGIÓN/IDEOLOGÍA. Por lo tanto *ética* es un término perteneciente a dos categorías y jerarquías diferentes.

Publicación, uso, evaluación y mantenimiento

El tesauro es el resultado de interpretar el HL teniendo en cuenta que sólo se publican las tuplas más frecuentes en caso de inconsistencias. Interpretar el HL es recorrer y representar visualmente, alfabéticamente, sistemáticamente y/o indexado el contenido del HL. Las tuplas "potencialmente inconsistentes", permanecen "invisibles" hasta que aumente su frecuencia con el uso. También pueden eliminarse si se considera necesario durante el proceso de evaluación.

Los errores que se han observado después de aplicar el método son de dos tipos: 1) escritura incorrecta de los términos, relaciones o categorías, por ejemplo errores mecanográficos, y 2) uso incorrecto del perfil de aplicación LOM, por ejemplo, taxones mal ordenados en un camino taxonómico. Estos errores sólo pueden ser detectados y corregidos por los usuarios que conocen cómo deben ser las clasificaciones de sus recursos.

La evaluación y mantenimiento del tesauro depende de la frecuencia con la que se utilice para explorar la colección de recursos. La estrategia que se sugiere es utilizar el tesauro para buscar recursos y, a la vez, detectar los errores e inconsistencias en el contenido y evaluar la eficacia del tesauro. Las modificaciones que se hagan en el tesauro deben servir, también, para modificar y corregir las clasificaciones de los metadatos de los recursos afectados, lo que contribuye a mejorar el sistema de indexación y la búsqueda de recursos.

En definitiva, cuanto más se utilice el tesauro y cuanto mayor tamaño tenga, más completo y menos errores tendrá Esta dependencia entre los recursos y el tesauro garantiza el mantenimiento y la evolución continúa del mismo a un coste razonable, siempre que se implementen mecanismos de actualización conjunta de los metadatos y el tesauro.

8.1.4. Resultados y discusión

Este caso práctico muestra el uso del método HL para llevar a cabo el proceso de especializar un tesauro general al dominio concreto de las colecciones de recursos educativos que manejan los docentes. La ventaja principal del método HL es que simplifica el proceso de especialización porque evita realizar las costosas y largas fases

de análisis, diseño conceptual y selección de fuentes, que son propias de los métodos de construcción generales (figura 7.1); con el método HL el usuario únicamente define e identifica las estructuras-t que ha construido anteriormente y que, en consecuencia, conoce perfectamente. Otro de los puntos fuertes del método HL es el uso del modelo HL, que según se ha demostrado, es capaz de representar no sólo la estructura del tesauro general, sino también las estructuras-t que, dinámicamente, se van integrando en el HL del tesauro para especializarlo. En consecuencia, es posible adaptar y mantener actualizado el contenido de los tesauros ya existentes sin dedicar un esfuerzo considerable a esta tarea aplicando el método HL conforme se añaden estructuras-t nuevas.

El método HL, sin embargo, tiene la desventaja de que es necesario trasladar el contenido del tesauro general de referencia a un sistema de almacenamiento basado en el modelo HL, por ejemplo una base de datos relacional con el esquema HL. Esta operación se puede llevar a cabo automáticamente pero, en cualquier caso, requiere el coste extra de tener que programar el cambio de modelo del tesauro general.

Respecto de la calidad de los resultados obtenidos, hemos constatado que debido al carácter inductivo y empírico del método HL, tenemos que tener en cuenta que:

i) la exhaustividad del tesauro depende del tipo de estructuras-t que se utilizan como fuente de datos para el tesauro. Cuando el tipo de estructuras-t está limitado, como en este caso, a un solo tipo de estructura semántica, TG/TE, el resultado es una menor riqueza relacional en el tesauro. No existen estructuras asociativas o de equivalencia para completar el tesauro. En cualquier caso, esto no siempre es una limitación, puesto que depende del objetivo del tesauro. En el caso del tesauro ETB-LRE (2008), el objetivo es que sirva básicamente de índice de las colecciones de recursos y no tanto de mapa conceptual del repositorio de recursos o de herramienta de recuperación. Por ello, en la revisión hecha por el grupo de trabajo de AENOR, no se consideraron las relaciones de equivalencia, aunque sí se mantuvieron las asociativas;

ii) la corrección del contenido depende de la corrección de las estructuras-t. Si las clasificaciones de los metadatos contienen errores, estos aparecen en el tesauro. Sin embargo, la estrategia de considerar sólo los términos y estructuras más frecuentes, minimiza el impacto de los posibles errores, tanto más cuanto mayor sea el número de estructuras-t insertadas. Además, la fase de revisión manual limpia el tesauro de errores y permite detectar y corregir los errores en las clasificaciones de los recursos afectados que, en caso de no corregirse serían difíciles de localizar. La fase de revisión y

corrección, por lo tanto, permite corregir no sólo el tesauro sino también los recursos que están clasificados inconsistentemente con respecto al tesauro y al resto de los recursos de la colección; y

iii) no existe un control del vocabulario definido explícitamente. Es responsabilidad del equipo de personas que crea las clasificaciones de los recursos educativos, los profesores, definir y mantener las reglas de control en la revisión y las relaciones de equivalencia entre términos si las consideran necesarias. Las variantes léxicas o variantes ortográficas de un término que aparecen en las clasificaciones LOM, se insertan en el tesauro, pero sin relacionar con sus formas preferidas o con sus formas alternativas. Supongamos que, como continuación del ejemplo (figura 8.3), se va a insertar otra clasificación LOM con el taxón *ámbitos de la moralidad* con el mismo significado que *ámbito moral*. El método propuesto simplemente incluiría este término en plural, pero no lo relaciona con su forma en singular. La razón es que el método no tiene como objetivo generar tesauros estándares de uso general. Al contrario, el propósito es generar tesauros personalizados y adaptados a un grupo de especialistas con unas necesidades concretas[6]. La consecuencia es que las reglas de control surgen con el uso, a partir las expresiones terminológicas más frecuentemente utilizadas por los usuarios-autores.

Finalmente, otras ventajas que aporta esta aproximación, frente a los métodos generales, son:

- la especialización, con un coste abordable para los equipos docentes universitarios, de un tesauro general de referencia general para adaptarlo a la gestión de colecciones específicas de recursos digitales;
- la reutilización del conocimiento (estructuras-t) de los materiales didácticos para ampliar el tesauro; y, en consecuencia
- el ajuste preciso al dominio de conocimiento;
- la evolución permanente del tesauro a un coste bajo; y
- un procedimiento de especialización que es compatible con una revisión manual más sistemática.

[6] Lo cual no implica que no se tengan en cuenta los estándares de construcción que aseguran la interoperabilidad y reusabilidad.

8.2. La reconstrucción, como tesauro, del índice temático de un museo virtual académico

Este segundo caso muestra la aplicación del modelo y método de HL para llevar a cabo un proceso de ingeniería inversa sobre el índice temático de un museo virtual académico de objetos virtuales. El propósito es recuperar las estructuras terminológicas en semántica libre, estructuras-t, que contiene el índice e interpretarlas conforme el modelo HL. Esta interpretación sirve para entender la estructura y significado del vocabulario que contiene el índice, y así, poder diseñar mecanismos que mejoren su eficacia. Si, además, se desea construir el tesauro de especialidad arqueología pre-colombina del museo que sistematiza la terminología del índice, se puede continuar aplicando las siguientes fases, revisión e inserción, del método inductivo HL. El tesauro obtenido mejorará la eficacia del sistema de acceso al museo porque hace explícitas las relaciones semánticas entre los términos, lo que ayuda al usuario a encontrar lo que se busca y a comprender lo que se encuentra. Además, el uso del modelo HL facilitará el mantenimiento y actualización posterior del tesauro, asegurando su evolución durante todo el ciclo de vida del museo.

8.2.1. Introducción

El museo CHASQUI[7] es un repositorio de objetos digitalizados y documentados, denominados Objetos Virtuales (OV) creados en el marco de varios proyectos de investigación[8]. Estos objetos provienen de los materiales arqueológicos y etnográficos[9] del Departamento de Historia de América II (Antropología de América)[10] de la Universidad Complutense de Madrid. El objetivo del museo es servir de apoyo a la docencia, al trabajo de los investigadores y a la creación de puntos de información, sitios Web, de contenido cultural.

El repositorio es un sistema informático de almacenamiento y gestión de objetos digitales documentados para crear, almacenar, clasificar, buscar, estudiar, analizar y reutilizar recursos educativos (Guinea, 2004; Sierra y Fernández-Valmayor, 2006). Ha sido construido por el Grupo de Ingeniería del Software e Inteligencia Artificial del

[7] http://macgalatea.sip.ucm.es/web/principal/principal.html
[8] Proyecto OdA-Virtual: Objetos de Aprendizaje en el campus Virtual (TIN2005-08788-C04-01), y proyectos anteriores: http://macgalatea.sip.ucm.es/web/infoProyecto/presentacion.php
[9] http://macgalatea.sip.ucm.es/web/infoProyecto/presentacion.php#materiales
[10] http://www.ucm.es/info/america2/

Departamento de Sistemas Informáticos y Programación[11] de la Facultad de Informática de la UCM.

Figura 8.6. Acceso clasificado al repositorio CHASQUI

Actualmente, el repositorio contiene más de 2700 objetos virtuales (figura 8.6), construidos con una metodología inductiva y cooperativa basada en el modelo de Objeto Virtual (OV) (Sierra, et. al, 2005). El modelo OV se compone de tres dimensiones: datos, metadatos, y recursos (figura 8.7). De esta forma para construir un OV se deben definir, por separado, las propiedades del objeto, su ficha de metadatos y el conjunto de archivos (texto, imágenes, …) que lo componen físicamente. Los metadatos de la ficha descriptiva se especifican utilizando un perfil de aplicación específico del estándar IEEE-LOM, denominado LOM-OV. En este perfil incluye un campo para clasificar los OV utilizando el vocabulario de especialidad de los profesores e investigadores que crean los OV. Las tres dimensiones del OV están siempre accesibles para su visualización y actualización, de forma que los usuarios del repositorio pueden ir creando, modificando y utilizando los OVs.

El acceso a los OV se realiza a través de los datos o de los metadatos. Las clasificaciones de los metadatos han servido, además, para construir un índice temático que aparece a la izquierda de la interfaz de acceso (figura 8.6). Este índice contiene siete categorías: Mat. Documental, Arqueología, Etnología, Reproducciones, Archivo gráfico

[11] http://www.ucm.es/info/dsip/

de Historia de América II, sin asignar. Cada categoría está estructurada en una o varias jerarquías. El índice sirve para clasificar, indexar y describir conceptualmente los OV. Es una herramienta de exploración del repositorio, que sin embargo, presenta algunos inconvenientes:

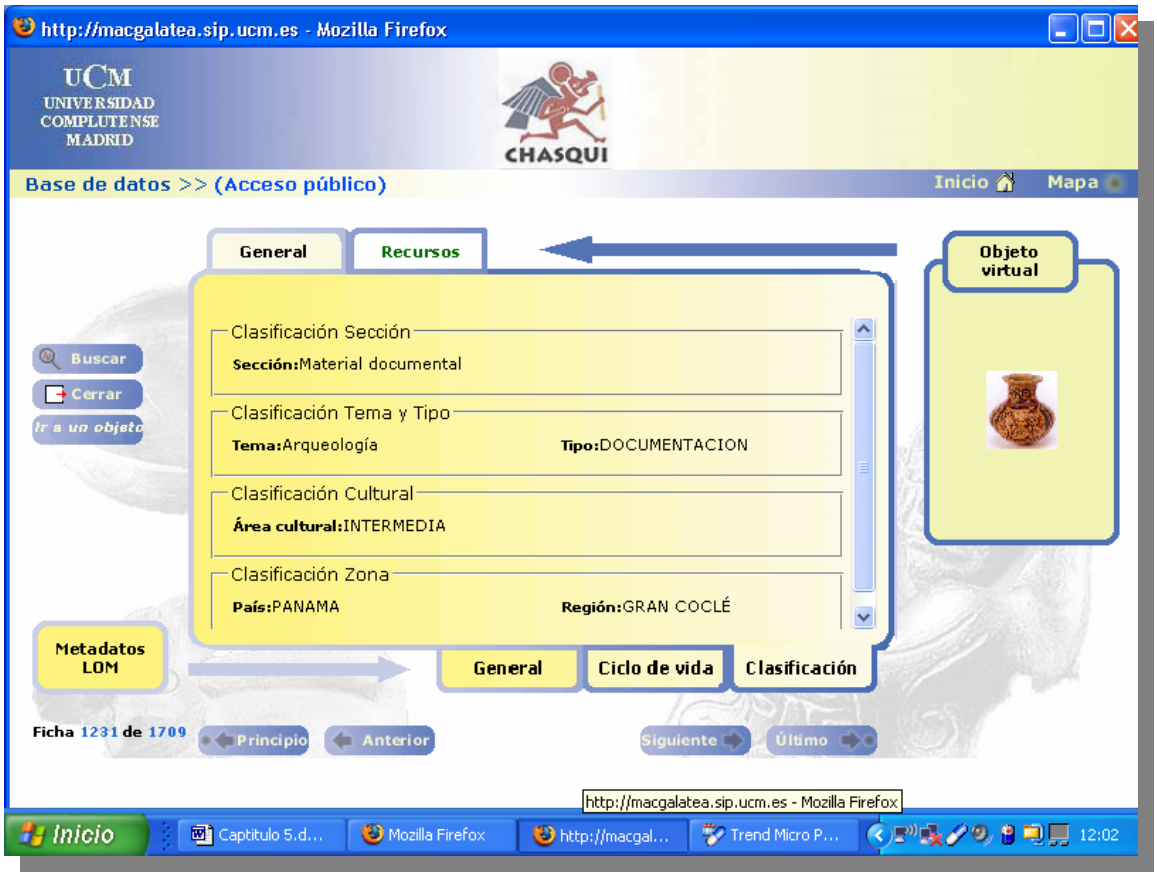


Figura 8.7. Presentación de un OV: datos (solapa superior General), metadatos (solapas inferiores) y recursos (solapa superior Recursos)

i) no están claras las relaciones semánticas entre los términos;

ii) es difícil ampliar la tipología de relaciones semánticas del índice, si fuera necesario; y

iii) es difícil reutilizar el contenido, ya que no existe un esquema de datos que sistematice el vocabulario del índice.

Estos inconvenientes dificultan la gestión, la comprensión de los términos, la explotación flexible del contenido y compartirlo o integrarlo en otros repositorios educativos.

Para resolver estos inconvenientes, aplicamos una aproximación basada en el modelo y método HL que reconvierte el índice en un tesauro específico del museo CHASQUI con un impacto mínimo en el resto del sistema CHASQUI.

8.2.2. El proceso de ingeniería inversa: la identificación, extracción e interpretación de estructuras-t

El proceso de análisis y comprensión del contenido del índice se realiza aplicando las tres primeras fases del método HL (capítulo 7, figura 7.6): 1) identificación y definición de estructuras-t, 2) extracción, y 3) análisis e interpretación.

Primera Fase: identificación y definición de estructuras-t

El índice temático fue construido con las clasificaciones de los metadatos de los OV del repositorio CHASQUI. Los profesores crearon las clasificaciones de los OV, utilizando su propio lenguaje de especialidad. Estas clasificaciones se escribieron dentro del campo *classification* del perfil de metadatos LOM-OV, siguiendo un convenio propio del equipo docente (figura 8.8).

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <manifest xmlns="http://www.imsglobal.org/xsd/imscp_v1p1" xmlns:imsmd="http://www.imsglobal.org/xsd/imsmd_v1p2"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" identifier="MANIFEST-1" xsi:schemaLocation="http://www.imsglobal.org/xsd/imscp_v1p1
  imscp_v1p1.xsd http://www.imsglobal.org/xsd/imsmd_v1p2 imsmd_v1p2p2.xsd">
  - <metadata>
    - <lom>
      + <general>
      + <lifecycle>
      - <classification>
        - <taxonpath>
          - <source>
              <langstring xml:lang="sp">Sección/País/Región/Yacimiento</langstring>
            </source>
          - <taxon>
            - <entry>
                <langstring xml:lang="sp">Atacames</langstring>
              </entry>
            </taxon>
          </taxonpath>
        - <taxonpath>
          - <source>
              <langstring xml:lang="sp">Sección/Tipo</langstring>
            </source>
          - <taxon>
            - <entry>
                <langstring xml:lang="sp">Práctica docente</langstring>
              </entry>
            </taxon>
          </taxonpath>
```

Figura 8.8 Extracto del código LOM-OV de una clasificación

Este convenio, definido de forma empírica, adapta el uso previsto del elemento 'classification' del modelo estándar IEEE-LOM, al modelo de clasificación de los objetos en el museo. El sistema de gestión de CHASQUI permite que las clasificaciones ya creadas puedan "reutilizarse" para clasificar otros OVs nuevos (figura 8.9). Esto significa que es posible calcular la frecuencia de uso de una clasificación, un dato importante para ayudar a validar la corrección de las estructuras semánticas de términos en el futuro tesauro HL.

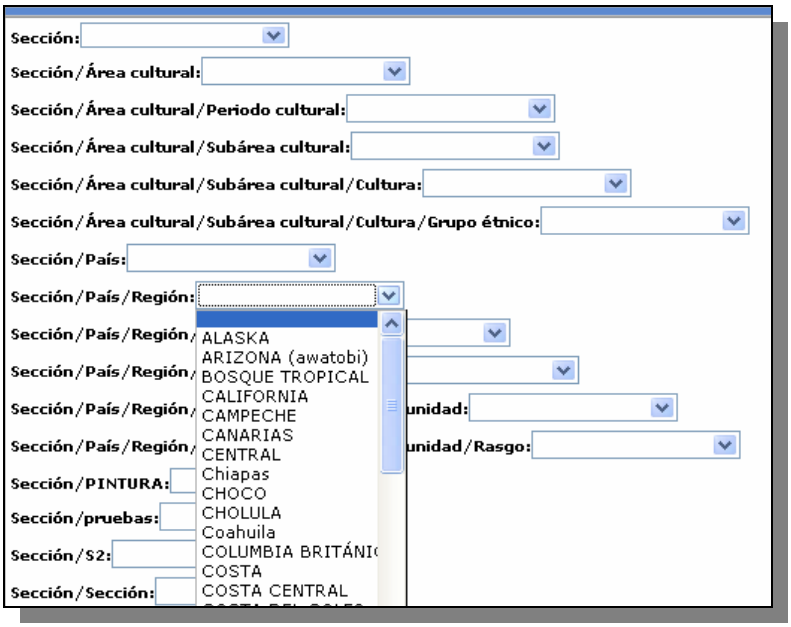


Figura 8.9. Reutilización de clasificaciones

Análisis de las clasificaciones del perfil LOM-OV

Las clasificaciones del perfil LOM-OV son la fuente de términos del índice temático, y también de las estructuras-t para construir el tesauro HL. Para identificar y extraer estas estructuras es necesario analizar el perfil de aplicación LOM-OV. Es un perfil sencillo, porque sólo se utilizan tres de los nueve campos principales del estándar IEEE-LOM: el campo general, ciclo de vida (cyclelife) y clasificación (classification) (figura 8.8)[12]. La etiqueta clasificación, que es la que interesa en este caso, contiene el conjunto de pares, secuencia de categorías y término, que sirven para clasificar e indexar un OV. Cada par está marcado por la etiqueta <taxonPath> y cada componente del par se marca de forma distintiva con la etiquetas <source> para la secuencia de categorías y <taxon><entry> para el término. Por ejemplo, el par:

```
<taxonPaht>
  <source>
        <langstring xml:lang="sp"> Sección/País/Región/Yacimiento </langstring>
  </source>
  <taxon>
    <entry>
        <langstring xml:lang="sp"> Atacames </langstring>
    </entry>
  </taxon>
</taxonPaht>
```

[12] La etiqueta general describe, a nivel global, el objeto incluyendo sus datos de catalogación del museo. La etiqueta ciclo de vida incluye los datos de gestión editorial del objeto –los autores, fecha de creación y modificación y estado de publicación.

Contiene primero la secuencia inclusiva de categorías ‘Sección’, ‘País’, ‘Región’ y ‘Yacimiento’, y utiliza como separador el carácter “/”. La primera categoría, ‘Sección’, es la más general (es la raíz) y la última, ‘Yacimiento’, es la más específica. El segundo componente del par es el término ‘Atacames’ (figura 8.10).

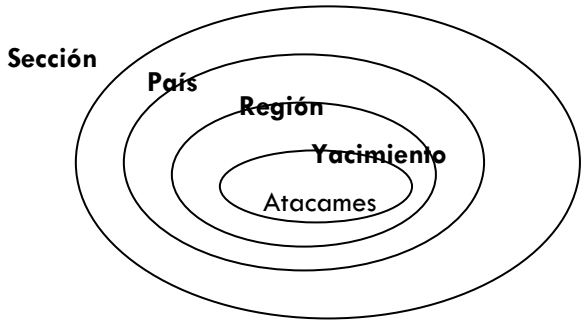


Figura 8.10. Interpretación gráfica de un par categorías/término de una clasificación LOM-OV

Todos los términos de una clasificación describen un mismo OV, por lo que están relacionados formando redes asociativas. Por ejemplo los términos CAPAYA, CONTEMPORÁNEO, ANDINA, ECUADOR, ESMERALDAS y ETNOLOGÍA son términos relacionados porque proceden de una misma clasificación que define un objeto (figura 8.11):

```
- <classification>
- <taxonpath>
    - <source>
      <langstring xml:lang="sp">Sección/Área cultural/Subárea cultural/Cultura</langstring>
      </source>
    - <taxon>
    - <entry>
      <langstring xml:lang="sp">CAYAPA</langstring>
      </entry>
      </taxon>
  </taxonpath>
- <taxonpath>
    - <source>
      <langstring xml:lang="sp">Sección/Área cultural/Periodo cultural</langstring>
      </source>
    - <taxon>
    - <entry>
      <langstring xml:lang="sp">CONTEMPORÁNEO</langstring>
      </entry>
      </taxon>
  </taxonpath>
- <taxonpath>
    - <source>
      <langstring xml:lang="sp">Sección/Área cultural</langstring>
      </source>
    - <taxon>
    - <entry>
      <langstring xml:lang="sp">ANDINA</langstring>
      </entry>
      </taxon>
  </taxonpath>
```

```
- <taxonpath>
    - <source>
      <langstring xml:lang="sp">Sección/País</langstring>
      </source>
    - <taxon>
    - <entry>
      <langstring xml:lang="sp">ECUADOR</langstring>
      </entry>
      </taxon>
      </taxonpath>
- <taxonpath>
    - <source>
      <langstring xml:lang="sp">Sección/Sección</langstring>
      </source>
    - <taxon>
    - <entry>
      <langstring xml:lang="sp">ETNOLOGIA</langstring>
      </entry>
      </taxon>
  </taxonpath>
- <taxonpath>
    - <source>
      <langstring xml:lang="sp">Sección/País/Región</langstring>
      </source>
    - <taxon>
    - <entry>
      <langstring xml:lang="sp">ESMERALDAS</langstring>
      </entry>
      </taxon>
  </taxonpath>
  </classification>
```

Figura 8.11. Clasificación de un OV del repositorio CHAQUI

Identificación y definición de las estructuras-t

Una clasificación LOM-OV, que se identifica con la etiqueta <classification>, contiene un conjunto de estructuras-t representadas por los caminos taxonómicos que se identifican la etiqueta <taxonpath>. A su vez, cada camino taxonómico contiene una secuencia de inclusión de categorías y un término perteneciente a ellas. Este par, secuencia de Categorías y Término, está marcado con los subelementos LOM del campo clasificación y con el metacarácter "/" conforme a las siguientes reglas (figuras 8.11 y 8.12):

- una clasificación incluye todos los caminos taxonómicos que describen a un OV, cada uno de ellos marcado con la etiqueta <taxonPath>;
- la etiqueta <source> se utiliza para definir las categorías y subcategorías en las que se incluye el término que contiene la etiqueta <taxon>. Ese término se incluye explícitamente sólo en la categoría más específica de la secuencia;
- el símbolo "/" se utiliza para indicar la relación de inclusión[13];
- la etiqueta <taxon> se utiliza para definir el término;

[13] A/B significa A⊂B.

- la etiqueta <entry> se utiliza para definir la forma ortográfica del término;
- la etiqueta <id> no se utiliza, pero podría utilizarse para definir el identificador único del término en un vocabulario. En el repositorio CHASQUI el <id> es asignado automáticamente por el sistema cuando se incorpora una nueva clasificación; y
- todos los términos de una clasificación definen un OV, por lo que forman una red semántica TR

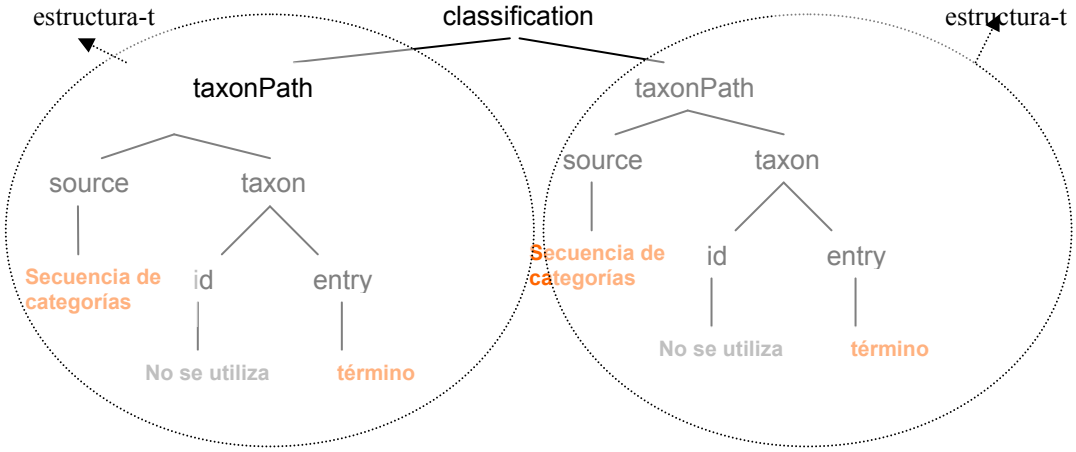


Figura 8.12. Definición de la estructura-t y sus marcas LOM-OV

Con este conocimiento de las clasificaciones de los OV del repositorio CHASQUI se define en la siguiente tabla de correspondencias (tabla 8.4):

MARCA/PATRÓN	**COMPONENTE**
<classification><taxonpath>	Estructura-t
<classification><taxonpath><source>CAT/CAT*	CAT+
/	I
<classification><taxonpath><taxon><entry>	TERM
Todos los TERM incluidos en una clasificación	TERM+
Son términos relacionados por definir un mismo OV	TR

Tabla 8.4. Definición del tipo de estructuras-t en CHAQUI

Segunda fase: extracción de estructuras-t

Una vez definido el tipo de las estructuras-t y su localización la operación siguiente es extraerlas de las fuentes, que en este caso son los archivos de metadatos de los OV de CHAQUI. Para ello, se aplica el mismo algoritmo que en el caso práctico anterior:

Entrada: fuente archivos metadata.xml de los OV
Para cada fuente:
1) Reconocer la estructura-t en el archivo de metadatos utilizando las marcas y patrones de la tabla de correspondencias, y
2) Extraer las estructuras-t
Fin bucle Para

Salida: conjunto de caminos taxonómicos (taxonPath)

La figura 8.13 muestra un extracto de los posibles caminos taxonómicos, referidos a lugares, que se obtienen como resultado de esta fase de extracción.

Figura 8.13. Estructuas-t extraídas del repositorio CHASQUI

Tercera fase: análisis e interpretación de las estructuras-t conforme el modelo HL

En esta fase se interpretan las estructuras-t extraídas en la fase anterior respecto del modelo HL (figuras 8.13 y 8.15). El resultado será el conjunto de tuplas HL que representan el significado de la clasificación de un OV conforme el modelo del profesor (tabla 8.4) y el modelo semántico HL[14] (figura 8.14). Estas tuplas van a representar un modelo de organización en el que todos los términos van a estar colocados en categorías, las categorías pueden contener subcategorías, y entre los términos que

[14] Recordamos que el significado de cada término es μ(término) = {conjunto de relaciones semánticas en las que participa}. El significado de cada categoría se calcula de la siguiente forma: μ(Categoría) = ⊗(∀i∈{1..k) (∪ (∀P∈πi(Categoría)) μ(P)). Si una categoría está formada por subcategorías ortoganes entre sí, su significado se calcula haciendo el producto cartesiano de los significados de las categorías ortogonales que lo constituyen. Si no, el significado de una categoría es la unión de los significados de las categorías y términos que las constituyen (ver 6.3.2 del capítulo 6).

clasifican un mismo objeto encontramos que existen redes semánticas, impícitas, de tipo asociativo (TR).

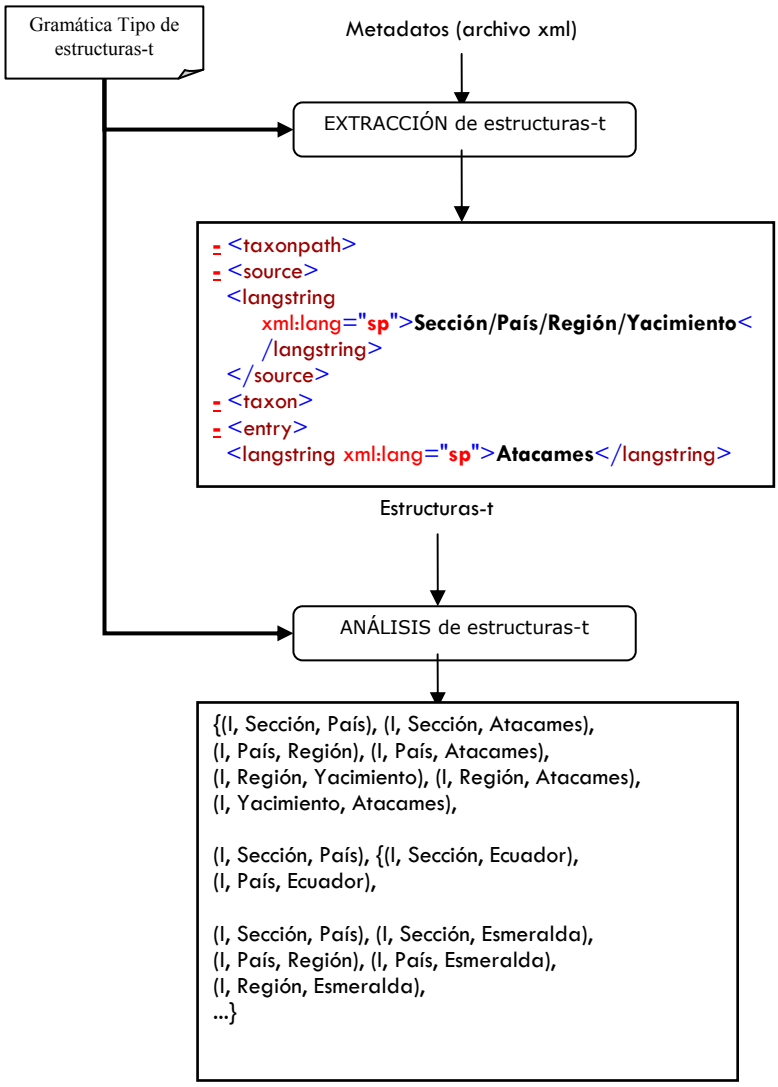


Figura 8.15. Visión conjunta de los procesos de extracción y análisis

El algoritmo de análisis e interpretación conceptual de las estructuras-t realiza las dos tareas:

1. Obtener las series semánticas de signos, interpretando cada estructura-t extraída conforme la tabla de correspondencias definida por el usuario, y
2. Interpretar las series semánticas conforme el modelo HL, utilizando el metalenguaje de definición de tipo de estructura-t. Para eso, se procesan modularmente cada tipo de serie (capítulo 7, sección 7.2.2).

Tomando como ejemplo las estructuras-t extraídas en la figura 8.13 se obtendrían, de la primera tarea, las series semánticas[15]:

CAT= Sección
CAT= País
CAT= Región
CAT= Yacimiento
I
TERM= Atacames

CAT= Sección
CAT= País
I
TERM= ECUADOR

CAT= Sección
CAT= País
CAT= Región
I
TERM= ESMERALDA

TERM=ECUADOR
TERM=ESMERALDA
TR

Después de la segunda tarea se obtiene la interpretación HL que, gráficamente, se representa en la figura 8.15:

{(I, Sección, País),
(I, País, Región),
(I, Región, Yacimiento),
(I, Sección, Atacames),
(I, País, Atacames),
(I, Región, Atacames),
(I, Yacimiento, Atacames)}

{(I, Sección, País),
(I, Sección, ECUADOR),
(I, País, ECUADOR)}

{(I, Sección, País),
(I, País, Región),
(I, Sección, ESMERALDA),
(I, País, ESMERALDA),
(I, Región, ESMERALDA)}

{(TR,ECUADOR,ESMERALDA)}

[15] Los términos ECUADOR Y ESPERALDA proceden de una misma clasificación (figura 8.11), por lo que forman entre sí una serie de términos relacionados.

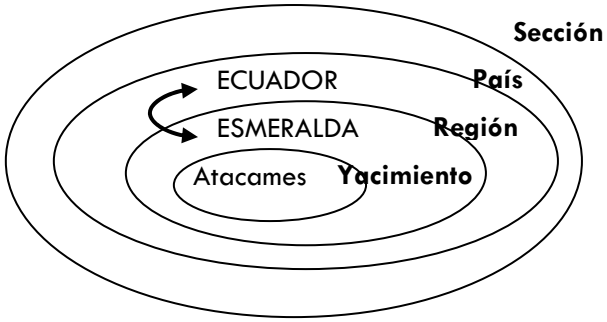


Figura 8.15. Interpretación gráfica de los componentes de las estructuras-t ejemplo[16]

Con este resultado, se obtiene una interpretación a nivel conceptual, independiente de la implementación final del HL. Se alcanza, por lo tanto, el primer objetivo de este caso práctico, que es analizar la naturaleza y funcionamiento del vocabulario que contiene el índice generado inductivamente en el repositorio CHASQUI utilizando el modelo y método HL. El resultado puede procesarse para su estudio, por ejemplo visualizando la estructura terminológica, o bien puede utilizarse para construir un tesauro que muestre explícitamente la forma que tiene el dominio conceptual que se ha creado con las clasificaciones de los OV.

8.2.3. El proceso de reconstrucción del índice como tesauro: inserción de las estructuras HL

En un proceso de ingeniería inversa que no tiene mucho sentido llevar a cabo la construcción del tesauro, sin embargo, para terminar de evaluar la viabilidad y posibilidades del método HL, se muestra el proceso de construcción con las siguientes fases, cuarta y quinta.

Cuarta fase: revisión y adecuación de la forma de los términos

Esta etapa permite corregir anomalías y errores que, como el ejemplo, en la figura 8. 13, introducen un uso no normalizado de mayúsculas y minúsculas en los signos del tesauro. Sin embargo, en este caso práctico no realizamos ninguna revisión y adecuación porque, en primer lugar, queríamos reproducir exactamente el índice del repositorio CHASQUI, incluyendo las irregularidades en las formas ortográficas[17] y en segundo lugar, porque se trata de un tesauro académico de explotación de tipo

[16] Las flechas indican relaciones TR entre los términos.
[17] Normalización de Términos que se han escrito de formas diferentes.

postcoordinado[18]. Por lo tanto, el tesauro se construye con un vocabulario que se considera ya normalizado.

Quinta fase: inserción de estructuras en el HL

Las estructuras de términos interpretadas en la fase 3 son conjuntos de tuplas que representan relaciones entre dos signos lingüísticos. Para insertarlas en un sistema de almacenamiento, base de datos, basado en HL es necesario ajustar estas representaciones a nivel conceptual con el nivel de implementación de datos. Si se utiliza el esquema de datos relacional (ver capítulo 7, sección 7.3.2) para la implementación de HL:

hl_micro(signo,tipo,frecuencia)
hl_macro(tipo_relacion,signo1,signo2, frecuencia)

Las inserciones tienen que tener en cuenta que:

i) hl_macro tiene una columna más, la frecuencia, para resolver las posibles incompatibilidades, registrando el número de veces que se utiliza cada tupla; y
ii) los términos que pertenecen a varias categorías relacionadas por inclusión, deben incluirse sólo en la categoría más específica para evitar la redundancia.

Para facilitar el registro de la frecuencia de uso de cada signo, término y categoría, y la relación semántica de las estructuras-t, la fase tercera, análisis, interpretación, y cuarta, inserción de las estructuras-t, se pueden realizar en una sola fase[19]. Para ello utilizamos el algoritmo general que, partiendo de las estructuras-t extraídas en la segunda fase, obtenga directamente las sentencias de inserción SQL en la base de datos relacional HL[20]:

Entrada: estructuras-t extraídas de las fuentes

Obtener las series semánticas;

Procesar Series de CAT, si existen;

Procesar la inclusión de los TERM en CAT, si existen;

Procesar Series de TERM, si existen;

Salida: sentencias de inserción SQL

Los algoritmos para procesar las series que hemos diseñado son los siguientes:

[18] Los términos se asignan a los objetos que se van a indexar como claves de recuperación, independientemente de sus relaciones gramaticales. Puede prescindirse de esta etapa porque es responsabilidad del software de búsqueda combinar los términos para no perder cobertura.

[19] Ya se comentó en la sección 7.3.2 que el análisis e interpretación y la inserción de estructuras-t podía hacerse directamente en una sola etapa cuando se tenía decidido el esquema de implementación de datos.

[20] El algoritmo describe las instrucciones de inserción en un lenguaje de alto nivel de abstracción. Sin embargo, en la traza del algoritmo que se presenta a continuación se especifican en el lenguaje SQL.

Procedimiento Procesar Series de CAT:

/* Primero se insertan las categorías en hl_micro y luego las relaciones de inclusión en hl_macro*/
Para cada serie de CAT **hacer**
Leer categorías;
categoría_fin:= primera categoría de la secuencia de Categorías;
insertar (categoría_fin, categoría) en hl_micro;
Mientras (Exista categoría_fin+1) **hacer**
Insertar (categoría_fin+1, categoría) en hl_micro;
Insertar (I, categoría_fin, categoría_fin+1) en hl_macro;
categoría_fin:= categoría_fin+1; /* avanza a la siguiente categoría */
finMientras
/*Termina cuando categoría_fin es la última categoría*/
finPara;

Procedimiento Procesar la inclusión de los TERM en CAT:

/*NOTA: construye la relación de inclusión de un término en la categoría más específica para evitar redundancia*/
Para cada serie de TERM **hacer**
Insertar (TERM, termino) en hl_micro sino se ha hecho antes;
Insertar (I, categoría_fin, TERM) en hl_macro;
finPara;

Procedimiento Procesar las Series de TERM:

REL:= relación semántica de la serie
Para i:= 1 hasta (longitud serie de TERM -1) **hacer**
Insertar (REL, TERMi, TERMi+1), en hl_macro;
finPara

El resultado es un conjunto de sentencias SQL que insertan en la base de datos HL y que se corresponden con las redes semánticas de las estructuras-t.

La tabla 8.5 muestra el funcionamiento de este algoritmo utilizando como entrada las estructuras-t de la figura 8.13:

SERIES SEMÁNTICAS	**SENTENCIAS SQL**
CAT= Sección CAT= País CAT= Región CAT= Yacimiento I TERM= Atacames	insert into hl_micro (signo, tipo) VALUES (‘Sección’, categoría);
	insert into hl_micro (signo, tipo) VALUES (‘País’, categoría); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’Sección’,’País’);
	insert into hl_micro (signo, tipo) VALUES (‘Región’, categoría); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’País’,’Región’);
	insert into hl_micro (signo, tipo) VALUES (‘Yacimiento’, categoría);

	insert into hl_macro (rel ,signo1 ,signo2) values ('I',’Región’,’Yacimiento’);
	insert into hl_micro (signo, tipo) VALUES (‘Atacames’, termino); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’Yacimiento’,’Atacames’);
CAT= Sección CAT= País I TERM= ECUADOR	insert into hl_micro (signo, tipo) VALUES (‘Sección’, categoría);
	insert into hl_micro (signo, tipo) VALUES (‘País’, categoría); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’Sección’,’País’);
	insert into hl_micro (signo, tipo) VALUES (‘ECUADOR’, termino); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’País’,’ECUADOR’);
CAT= Sección CAT= País CAT= Región I TERM= ESMERALDA	insert into hl_micro (signo, tipo) VALUES (‘Sección’, categoría);
	insert into hl_micro (signo, tipo) VALUES (‘País’, categoría); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’Sección’,’País’);
	insert into hl_micro (signo, tipo) VALUES (‘Región’, categoría); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’País’,’Región’);
	insert into hl_micro (signo, tipo) VALUES (‘ESMERALDA’, termino); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’Región’,’ESMERALDA’);
TERM=ECUADOR TERM=ESMERALDA TR	insert into hl_macro (rel ,signo1 ,signo2) values ('TR','ECUADOR’,’ESMERALDA’);

Tabla 8.5. Traza del algoritmo de inserción[21]

La secuencia de instrucciones SQL generada en este ejemplo es:

```
insert into hl_micro (signo, tipo) VALUES (‘Sección’, categoría);
```

[21] No considera la actualización de la frecuencia ni la asignación de un identificador, porque se realiza automáticamente, dependiendo de los valores ya almacenados en la base de datos.

```
insert into hl_micro (signo, tipo) VALUES (‘País’, categoría);
insert into hl_macro (rel ,signo1 ,signo2) values ('I','Sección','País');
insert into hl_micro (signo, tipo) VALUES (‘Región’, categoría);
insert into hl_macro (rel ,signo1 ,signo2) values ('I','País','Región');
insert into hl_micro (signo, tipo) VALUES (‘Yacimiento’, categoría);
insert into hl_macro (rel ,signo1 ,signo2) values ('I','Región','Yacimiento');
insert into hl_micro (signo, tipo) VALUES (‘Atacames’, termino);
insert into hl_macro (rel ,signo1 ,signo2) values ('I','Yacimiento','Atacames');
insert into hl_micro (signo, tipo) VALUES (‘Sección’, categoría);
insert into hl_micro (signo, tipo) VALUES (‘País’, categoría);
insert into hl_macro (rel ,signo1 ,signo2) values ('I','Sección','País');
insert into hl_micro (signo, tipo) VALUES (‘ECUADOR’, termino);
insert into hl_macro (rel ,signo1 ,signo2) values ('I','País','Ecuador');
insert into hl_micro (signo, tipo) VALUES (‘Sección’, categoría);
insert into hl_micro (signo, tipo) VALUES (‘País’, categoría);
insert into hl_macro (rel ,signo1 ,signo2) values ('I','Sección','País');
insert into hl_micro (signo, tipo) VALUES (‘Región’, categoría);
insert into hl_macro (rel ,signo1 ,signo2) values ('I','País','Región');
insert into hl_micro (signo, tipo) VALUES (‘ESMERALDA’, termino);
insert into hl_macro (rel ,signo1 ,signo2) values ('I','Región','ESMERALDA');
insert into hl_macro (rel ,signo1 ,signo2) values ('TR','ECUADOR','ESMERALDA');
```

La ejecución de las sentencias agrega y/o confirma las siguientes filas en las tablas hl_micro y hl_macro[22] (tablas 8.6 y 8.7):

id	signo	tipo	freq
#	Sección	categoría	# +3
#	País	categoría	# +3
#	Región	categoría	# +1
#	Yacimiento	categoría	# +1
#	Atacames	término	# +1
#	ECUADOR	término	# +1
#	ESMERALDA	término	# +1

Tabla 8.6 Inserciones en la relación hl_micro

Tipo_relación	id_rel	signo1	signo2	inconsistente	freq
I	#	Sección	País	null	# +3
I	#	País	Región	null	# +2
I	#	Región	Yacimiento	null	# +1
I	#	Yacimiento	Atacames	null	# +1
I	#	País	ECUADOR	null	# +1
I	#	Región	ESMERALDA	null	# +1
TR	#	Ecuador	ESMERALDA	null	# +1

Tabla 8.7. Inserciones en la relación hl_macro

[22] Los identificadores de las relaciones (id_rel) y la frecuencia de uso no se explicitan porque dependen de los valores de la base de datos en el momento de la combinación. Para indicar estos valores anteriores que son desconocidos se utiliza el símbolo #.

Integración del tesauro en el repositorio CHASQUI

Como el propósito del tesauro es indexar, clasificar y definir los OV del repositorio CHASQUI, cada OV debe estar asociado a uno o varios términos del tesauro (figuras 8.7 y 8.16).

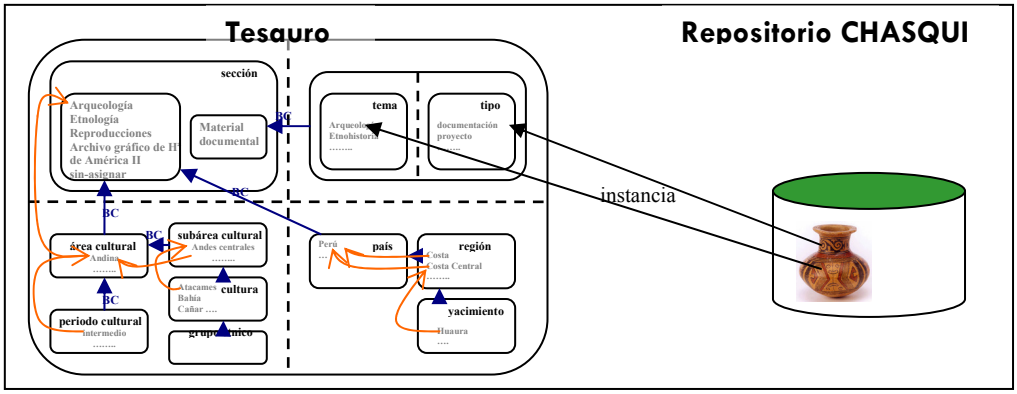


Figura 8.16. Relación del tesauro con un OV

Para integrar el tesauro HL en el sistema CHASQUI también es necesario ampliar la base de datos HL del tesauro, añadiendo una tabla de conexión entre los términos del tesauro HL y los OV del sistema CHASQUI que los utilizan en sus clasificaciones (figura 8.17). Los datos de la tabla de conexión –el identificador de término y el identificador del OV- se obtienen también de los metadatos de los OV.

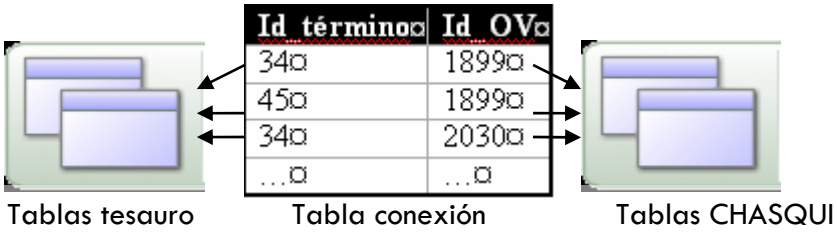


Figura 8.17. Integración del tesauro en el sistema CHASQUI

Etapa sexta: uso, validación y mantenimiento

El tesauro tiene como objetivo documentar y encontrar OV del repositorio. La documentación, como ya se ha comentado, la realizan los profesores utilizando el esquema de metadatos LOM-OV. El tesauro proporciona los términos para clasificar en los metadatos los OV (figura 8.9).

Las búsquedas y selección de OV con el tesauro, están basadas en la navegación, explorando el HL que modela el contenido del tesauro. La recuperación basada en lenguaje natural o por palabras no es rentable con este tesauro porque no se han incluido términos equivalentes que sirvan para expandir la consulta del usuario.

La estrategia de validación y mantenimiento del método HL se ve favorecida por la metodología de construcción de los OV del repositorio CHASQUI de carácter inductivo y colaborativo, porque permite:

1. continuar actualizando el tesauro con nuevos términos y/o categorías, en caso de que los términos del tesauro no sean adecuados para documentar nuevos OV; y
2. asegurar la validación y mantenimiento del tesauro. Los profesores, además, de crear las clasificaciones de los OV, que son las estructuras-t para construir el tesauro, las utilizan para buscar los OV. Este doble papel creador-usuario, favorece la detección y corrección de errores de forma permanente e integrada con el uso del tesauro, que será tanto más correcto cuanto más se utilice el tesauro. Los errores que han corregido son de tipo ortográfico, y de reorganización estructural. En la figura 8.18 se puede observar un ejemplo del primero. Las imágenes muestran el contenido de la categoría "Área cultural" (que es subcategoría de la categoría "Archivo Gráfico de Hª de América II") en dos tiempos diferentes. La imagen de la izquierda está tomada en el curso 2004/05 y la imagen de la derecha en el 2005/06.

Figura 8.18. Control y corrección de los términos TIAHUANACO y SIOUX.

Como se observa en la imagen de la izquierda existe un error ortográfico en el término "TIAHUANACO. (1)"[23] , detectado y corregido por los miembros del equipo docente en la imagen de la derecha por "TIAHUANACO C. (1)". También se observa una inconsistencia en el término SIOUX. A la izquierda, por alguna razón, el término está duplicado[24] y en una de las apariciones está asociado a cinco OV mientras que en la otra a un OV. A la derecha se observa la corrección, puesto que existe un único término SIOUX con seis OV asociados.

[23] Los términos tienen a su derecha entre paréntesis el número de objetos virtuales asociados.
[24] En la base de datos aparecía con dos identificadores diferentes.

En la figura 8.19 se muestra un ejemplo de modificación del esquema del tesauro. A la izquierda se representan las categorías "tema y tipo", que son dos subcategorías ortogonales. A la derecha se observa un cambio en la estructura taxonómica de las categorías, a la que se le añade la categoría "tema y tipo" que es el producto escalar de las categorías tema y tipo.

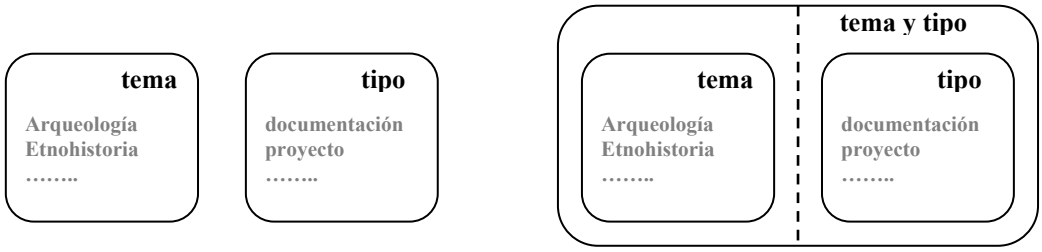


Figura 8.19. Ejemplo de modificación estructural en el tesauro

Esta modificación se puede realizar sin inconvenientes con el modelo HL, puesto que simplemente consiste en añadir, a la base de datos, dos relaciones más de partición entre las categorías "tema y tipo": (π, tema y tipo, tema), (π, tema y tipo, tipo).

8.2.4. Resultados

La aplicación del método HL para la sistematización del índice del repositorio CHASQUI conforme al modelo HL ha permitido obtener el HL que representa el contenido terminológico del índice. Este contenido se organiza en dos estructuras superpuestas: un conjunto de jerarquías disjuntas de inclusión de categorías (figuras 8.20 y 8.21) combinada con un conjunto de redes de signos relacionados semánticamente por asociación (figura 8.22).

El tesauro, representado por la categoría principal "root", está organizado en cinco facetas principales (figuras 8.7 y 8.20) que los profesores han denominado 'clasificación Sección', 'clasificación Tema y Tipo', 'clasificación Cultural' y 'clasificación Zona'. La faceta 'clasificación Sección' contiene una única categoría que es 'Sección'; la faceta 'clasificación Tema y Tipo' contiene dos subfacetas (categorías disjuntas) 'Tema y Tipo'; la faceta 'clasificación Cultural' contiene dos jerarquías de categorías, la primera tiene como categoría raíz 'Área cultural' y la segunda 'Periodo cultural'; la faceta 'clasificación Zona' contiene una jerarquía de categorías cuya raíz es 'País'. La figura 8.21 muestra la estructura jerárquica de categorías del tesauro.

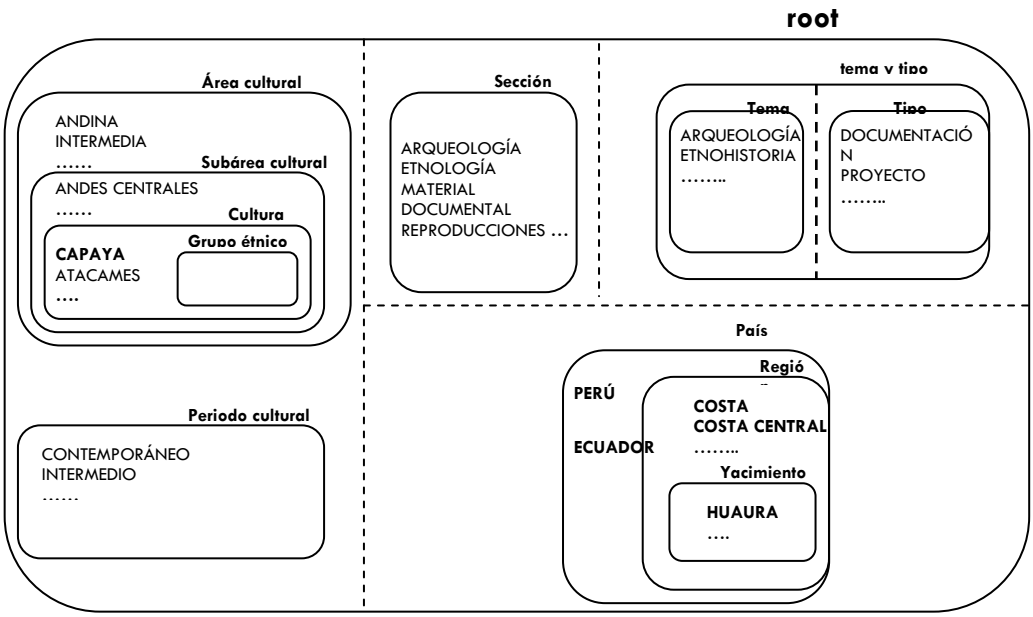


Figura 8.20. Estructura de las categorías del tesauro

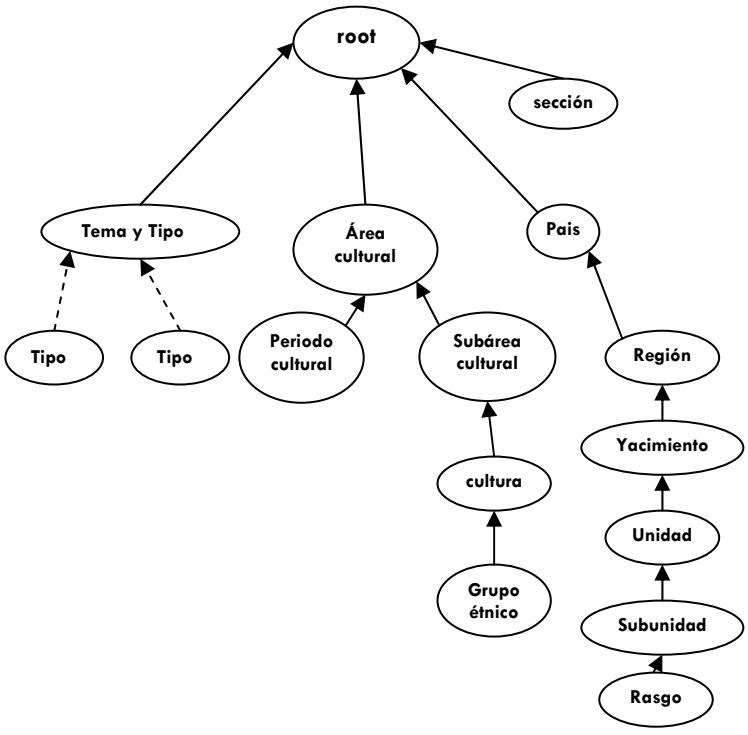


Figura 8.21. Estructura jerárquica de inclusión de las categorías

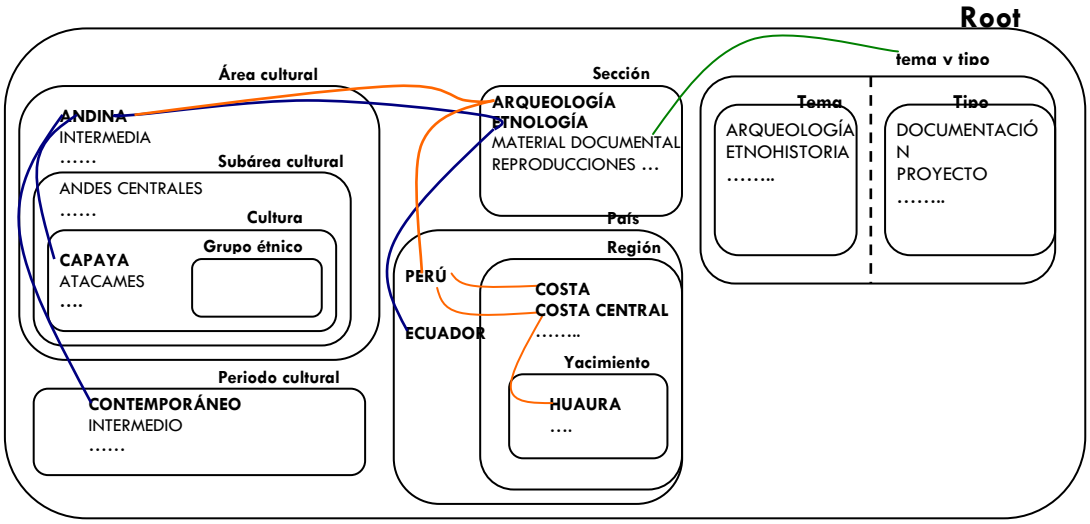


Figura 8.22. Redes de signos, términos y categorías, relacionados

Esta estructura de categorías clasifica los términos del tesauro. Los términos están asociados a los OV que representan. En la interfaz de búsqueda se muestran con un número entre paréntesis que indica que número de OV asociados (figura 8.23).

Figura 8.23. Términos relacionados con "Arqueología"

Los términos, además de estar clasificados en categorías, están organizados en redes semánticas TR. Esta relación crea una red TR de términos que definen un OV. Cada red, por lo tanto, es la representación del significado de, al menos, a un OV.

Las relaciones se establecen entre términos y entre términos y categorías y categorías entre sí. Por ejemplo, la figura 8.23, representa la relación entre el término 'Material documental' (en la categoría 'Sección') y la categoría 'Tema y Tipo'. Esto significa que todos los términos de la categoría están relacionados con 'Material documental'. En las jerarquías de 'Área cultural' y 'País' ocurre que los términos de las categorías superiores están relacionados con un grupo términos de la(s) categoría(s)

inmediatamente inferior(es), formando clases de equivalencia (figura 8.24). Esta relación entre términos de una misma categorías relacionados con un mismo término de la categoría superior podría considerarse una relación de generalización/especialización (TG/TE). En cualquier caso, las redes TR son útiles para precisar y orientar al usuario en su búsqueda, de forma que si se selecciona un término o categoría, éste se expande mostrando sólo los términos o categorías con los que está relacionado (figura 8.23).

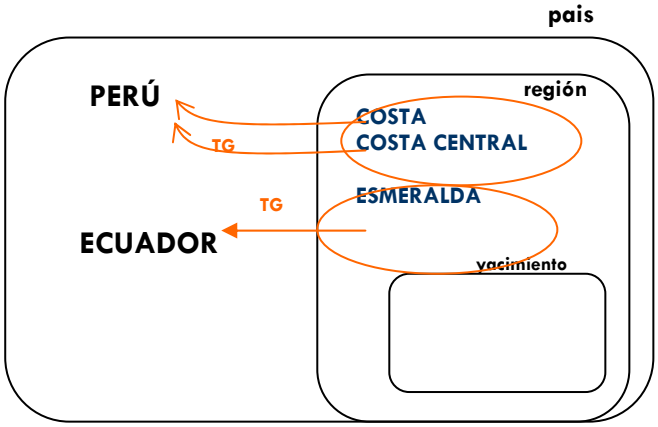


Figura 8.24. Clases de equivalencia de términos específicos

Formalmente se obtiene un HL, con los siguientes componentes:

$$HL = (S, \sigma, \pi, R)$$

- El conjunto de signos que lo constituyen, 16 categorías y aproximadamente 250 términos: Signos = Categorías ∪ Términos;
- La función σ, que establece una clasificación de los términos en las categorías y las categorías en subcategorías. Por ejemplo[25]:

  σ(Sección) = {ARQUEOLOGÍA, ETNOLOGÍA, REPRODUCCIONES, MATERIAL DOCUMENTAL, ARCHIVO GRÁFICO DE HISTORIA DE AMÉRICA II, ~sin asignar}
  σ(Área cultural) = {Subárea cultural, ANDINA, INTERMEDIA, CENTROAMÉRICA, AMAZÓNICA, CIRCUMCARIBE, …}
  σ(Subárea cultural) = { Cultura, ANDES CENTRALES, ANDES SEPTENTRIONALES, EXTREMO NORTE, …}

- La función π, que define las facetas principales del tesauro (Root):

  Root = Sección x Tipo y Tema x Área Cultural x País
  π_1(Root) = Sección, π_2(Root) = Tipo y Tema, π_3(Root) = Cultura, π_4(Root) = Zona
  π_1(Tema y Tipo) = Tema
  π_2(Tema y Tipo) = Tipo

[25] Los términos se escriben en mayúsculas y las categorías en minúsculas y la primera letra en mayúsculas.

- Las Relaciones, R, que define el conjunto de relaciones semánticas del tesauro, que en este caso son sólo relaciones de tipo TR:

  R = {(TR, MATERIAL DOCUMENTAL, Tema y Tipo), (TR, ANDINA, ANDES CENTRALES), (TR, ANDINA, ANDES CENTRO-SUR), (TR, ANDINA, ANDES SEPTENTRIONALES), (TR, ANDINA, EXTREMO NORTE),…, (TR, ECUADOR, ESMERALDAS), (TR, PERÚ, HUAURA),…}

8.2.5. Discusión

El objetivo de esta aplicación era demostrar que el modelo y método HL podía contribuir a mejorar la comprensión, gestión e integración de los índices de términos creados para el acceso al contenido de sistemas de información y de objetos digitalizados de forma empírica, inductiva y cooperativa por grupos de profesores y estudiantes. Los resultados conducen a las siguientes conclusiones:

(1) la aplicación del modelo HL permite sistematizar el contenido del índice de clasificación de los OV del museo académico CHASQUI en un esquema conceptual que surge inductivamente. Este esquema es una representación más cercana a la estructura conceptual que tienen los profesores del dominio que la que se obtenía con el índice original de CHASQUI o que la que se tiene con tesauros de carácter general, por lo que se puede explotar para:
   a. la exploración del repositorio de OV;
   b. ayudar al usuario poco experto, por ejemplo, el estudiante, a entender y aprender el lenguaje de especialidad que expresa la estructura conceptual de este dominio de conocimiento;
   c. ayudar al experto a corregir y refinar la estructura terminológica del tesauro que representa la estructura conceptual del dominio de conocimiento al que pertenecen los OV;
   d. ayudar al experto a localizar y reutilizar los OV que contiene la colección de OV y a documentar con el tesauro nuevos OV;
   e. proporcionar un marco conceptual coherente y expresado con una terminología familiar del museo CHASQUI; y
   f. procesar la colección de objetos con un enfoque más semántico que permita, por ejemplo, la integración y difusión del repositorio en repositorios federados, en Internet o aplicar algoritmos más inteligentes de recuperación del contenido del repositorio basados en el tesauro;

(2) la aplicación del modelo HL permite hacer explícitas las estructuras semánticas contenidas implícitamente en el índice, incluso aquellas que no son evidentes como por ejemplo, las redes de términos TR que no se habían podido detectar en el índice original;

(3) el método HL permite, como en el caso 1, construir el tesauro de explotación del repositorio CHASQUI, a partir de las clasificaciones de los metadatos LOM de los recursos digitalizados;

(4) el método HL puede seguir aplicándose a las clasificaciones de los metadatos de los nuevos recursos que se vayan creando para mantener actualizado el tesauro;

(5) sin embargo, al comparar este caso con el 1 se pone en evidencia la necesidad de abordar de forma manual, con la intervención de los profesores autores de las estructuras-t, la primera fase del método, la identificación y definición de estas estructuras-t. Aunque las estructuras-t del caso práctico anterior y de éste caso provienen de un mismo tipo de fuente estándar, que son las clasificaciones de los metadatos LOM, existe una falta de uniformidad en el uso y la interpretación de estas fuentes por parte de los equipos docentes. Así, aunque se trate de un mismo tipo de fuente de estructuras-t, la identificación de las estructuras-t tiene que hacerse, inicialmente, con ayuda del usuario quien, utilizando una muestra de clasificaciones del repositorio particular, debe indicar cómo las interpreta identificando los elementos básicos de las estructuras: términos, categorías, relaciones I, TG-TE, TR y USE-USEPARA;

(6) el método HL utiliza las clasificaciones que han sido cuidadosamente creadas y/o seleccionadas por los profesores e investigadores especialistas para documentar los objetos del área de conocimiento de arqueología pre-colombina. Esto asegura (a) la calidad los términos y estructuras semánticas de términos para la construcción del tesauro de arqueología precolombina del museo; (b) la familiaridad del usuario con la terminología; y (c) la precisión del tesauro para describir la colección de objetos del museo. Existe, sin embargo un inconveniente y es que los términos no están normalizados, por lo que el control del vocabulario debe llevarse a cabo manualmente con posterioridad, una vez construido el tesauro o bien en la etapa intermedia 4 de revisión y adecuación de términos (ver sección 8.2.3). La experiencia con CHASQUI, sin embargo, demuestra que da buenos resultados realizarla posteriormente, durante el uso del tesauro y no es excesivamente costoso

realizar este tipo de correcciones a la vez que se utiliza el tesauro como se ha descrito en la etapa 6 (ver sección 8.2.3);

(7) el tratamiento de inconsistencias requiere que las estructuras-t tengan una cierta calidad. Si las estructuras son dispares o de baja calidad el HL resultado no va a ser coherente. Esto ocurre cuando los autores de las estructuras-t no son expertos[26] y/o no disponen de principios o criterios mínimos para realizar las clasificaciones de los OV;

(8) como en el caso práctico 1, corroboramos que la aplicación del método HL reduce los costes de producción del tesauro porque reutilizamos las estructuras-t creadas por los profesores para sus recursos didácticos, no es necesario la selección y recolección de fuentes de términos, ni el análisis del dominio y el diseño del esquema de datos del tesauro; y

(9) el modelo HL se implementa con el mismo modelo de datos que el repositorio, el modelo relacional, lo que permite la integración del tesauro con un mínimo impacto en el sistema.

8.3. La creación de un tesauro para el *glosario explicativo e-derecho*

Este caso práctico presenta la aplicación del modelo y método HL en la construcción de un tesauro para la exploración del contenido de un glosario explicativo del ámbito del derecho y la propiedad intelectual en Internet. El tesauro reproduce las estructuras de términos relacionados que los autores del glosario habían incluido en el material original. El objetivo es que sirva para ayudar al usuario a localizar y entender la información que necesita explorando los términos más cercanos al concepto o conceptos que necesitan consultar en un área nueva, relativa al derecho e Internet.

8.3.1. Introducción

El *glosario e-derecho* surge en la actividad universitaria, por la necesidad urgente de resolver eficazmente los problemas e incertidumbres advertidos en la aplicación del Derecho tradicional común y de la propiedad intelectual a las obras y creaciones intelectuales desarrolladas o difundidas por los profesores, investigadores y estudiantes en entornos electrónicos de formación universitaria, campus virtuales. Este glosario

[26] Por ejemplo, cuando son los estudiantes los que escriben las clasificaciones sin la guía de un profesor.

electrónico es un modelo empírico que puede ser utilizado en múltiples ámbitos de la sociedad. En él se recoge y sistematiza el trabajo colaborativo y multidisciplinar de numerosos especialistas. Mediante 345 términos se ofrecen soluciones y alternativas para la regulación coordinada de los campos jurídicos implicados en el marco científico y tecnológico de la sociedad de la información y del conocimiento.

El glosario se ha construido en el marco de un Proyecto de Innovación y Mejora de la Calidad de la Docencia (PIMCD-66/2007-2008) financiado por la Universidad Complutense de Madrid[27]. Actualmente se utiliza en diferentes cursos virtuales del Campus Virtual UCM para el aprendizaje de los nuevos conceptos y términos que cubre.

Figura 8.24. Interfaz Web del Glosario y del Tesauro E-Derecho

El material original del glosario ha sido sistematizado con un modelo de contenido nuevo, extraído inductivamente del material original que estaba elaborado sin criterios lexicográficos (Flores, et. al, 2009). El modelo de contenido consta de 31 elementos estructurales organizados jerárquicamente en cinco niveles y linealmente en 4 dimensiones (figuras 8.24 y 8.25). Cada dimensión es una de forma independiente de entender y de ver la información, y para lograr una comprensión completa es necesario considerar todas las dimensiones: Significado, Enciclopedia[28], Ámbito jurídico[29] y Tesauro.

[27] Está disponible en la dirección http://www.ucm.es/info/contratos

[28] La dimensión Enciclopedia recoge la información que puede ayudar al lector a entender mejor el significado, por ejemplo, explicaciones y ejemplos.

Para consultar el glosario el usuario dispone de cuatro formas de acceso:

- la lista alfabética de términos, para buscar seleccionando la letra de interés y/o los términos en el listado de la derecha;
- filtrar/buscar todos los términos que contengan alguna palabra, o, los términos que cumplan algún criterio respecto de la forma gráfica del término[30];
- buscar en el contenido, utilizando como campos de búsqueda los elementos estructurales definidos en el modelo del glosario[31]; y
- navegando en el tesauro, seleccionando términos y dentro de cada término, seleccionando, con las pestañas superiores, otras dimensiones de la información del glosario.

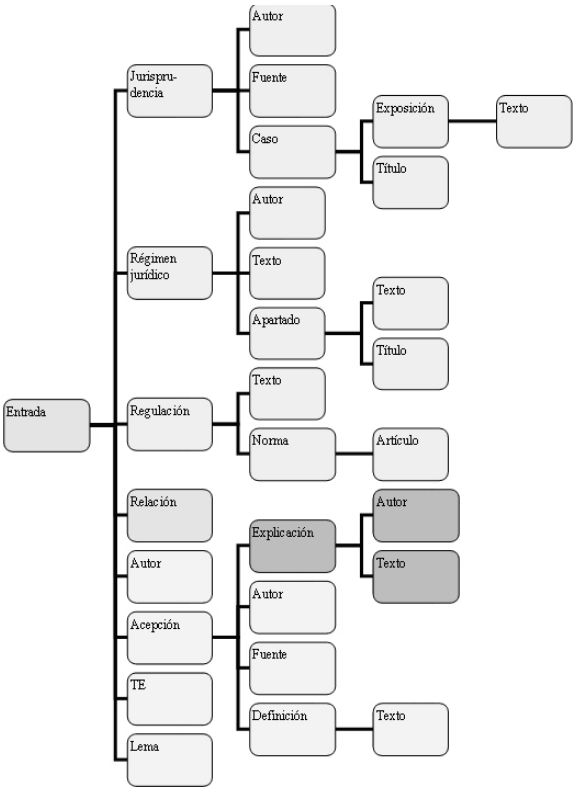


Figura 8.25. Estructura jerárquica del contenido del glosario e-derecho

El Tesauro del *glosario e-derecho* presenta los términos del glosario organizados en redes de relaciones semánticas TR, TG y USE y en tres categorías o áreas temáticas: Técnológica, Jurídica civil y Jurídica penal. Esta información se encuentra incluida en el

[29] La dimensión Ámbito jurídico recoge toda la información de carácter jurídico que suplementa y complementa el significado del término: regulación, jurisprudencia, explicaciones, ejemplos, etc.
[30] Por ejemplo, que contengan una determinada palabra, que la excluyan, que comiencen por una letra o forma ortográfica precisa.
[31] Por ejemplo, cuáles son los términos que contienen una norma determinada, o que se refieren a un ámbito concreto (Tecnológico, Jurídico Civil, Jurídico Penal) o los que han sido creados por un determinado autor.

contenido del glosario, concretamente en el elemento estructural "Relación" del glosario, que explicita las relaciones semánticas del lema de una entrada y otros términos del glosario, y en el elemento Acepción, que contiene la categoría temática a la que pertenece el término lema (figuras 8.26 y 8.27). La presentación del contenido del tesauro es alfabética-hipertextual, en consonancia con el sistema de acceso al glosario. Para utilizar el tesauro se selecciona, desde una entrada del glosario, la pestaña "Tesauro", que muestra el conjunto de términos directamente relacionados con el lema de la entrada (figura 7.34). Cada término es un hipervínculo que permite navegar por el tesauro y el glosario.

El objetivo de la dimensión Tesauro es proporcionar un mapa terminológico-conceptual con el conjunto de términos del glosario para: 1) facilitar, al usuario experto, la localización de la información que necesita, y 2) ayudar al usuario poco experto a entender y aprender el lenguaje, signos, significados y contexto de uso, del ámbito del derecho electrónico.

Respecto a la búsqueda de información, el tesauro complementa los sistemas de acceso al glosario, proporcionando un mecanismo para localizar el término o términos más cercanos semánticamente a uno dado. De esta forma, el usuario puede, a partir de un concepto que desea consular, elaborar toda la información relativa a él. El procedimiento es utilizar los mecanismos de búsqueda para localizar uno o varios términos relacionados y, a partir de ellos, navegar en el tesauro para ir encontrando y completando la información.

Respecto al uso del tesauro y del glosario como recursos didácticos, se utilizaron en el curso 2008-09, en dos experiencias para el aprendizaje de los conceptos y términos del ámbito de e-derecho. El tesauro se usó para la exploración, "guiada" por las relaciones semánticas, de los términos y contenido del glosario[32]. Las dos experiencias se llevaron a cabo en espacios virtuales del Campus Virtual UCM[33].

Para la creación del glosario se utilizó una herramienta de edición de diccionarios basada en XML, -TschwaneLex (http://tshwanedje.com/)-. Este tipo de herramientas tiene importantes ventajas para la construcción de una obra en la que participan múltiples autores y se abordan múltiples disciplinas: interfaz cercana al usuario para

[32] El primero de los experimentos proponía la lectura reflexiva y comentada de páginas Web del ámbito del e-derecho, en la que era necesario relacionar los términos que aparecían en el texto. En el segundo los estudiantes elaboraron un corpus de ejemplos reales de aplicación de varios conceptos expresados con términos del glosario e-derecho.

[33] No se han publicado todavía los resultados.

introducir y visualizar la información, el control de la consistencia de la información, el control de versiones y la exportación del contenido en diferentes formatos. Sin embargo, presenta varios inconvenientes para construir y mantener obras con contenidos altamente relacionados, como el caso del tesauro. Destacamos, como más importantes que 1) no es posible visualizar y gestionar la macroestructura de la obra; 2) no permite aplicar cualquier modelo de datos, ya que es necesario adaptar un modelo propio que, en el caso de información relacionada, genera redundancia; y 3) no es posible gestionar y mantener el tesauro sin que se vea afectado globalmente el glosario, los cambios que afectan únicamente al tesauro suponen cambiar e instalar una nueva versión del glosario.

Estos inconvenientes afectan especialmente a la eficacia del tesauro como instrumento de explotación del glosario, especialmente en su aspecto didáctico. Por ello, se optó por separarlo del glosario, y construirlo a partir del contenido del glosario, como un sistema léxico independiente, pero asociado, al glosario. La construcción se realizó siguiendo el método inductivo HL. Esta experiencia es la que se presenta en este caso práctico.

8.3.2. Análisis del tesauro del *glosario e-derecho*

Los autores del material original del *glosario e-derecho* consideraron importante señalar, dentro del contenido de una entrada, los términos que estaban relacionados semánticamente con el lema de esa entrada (figura 8.26).

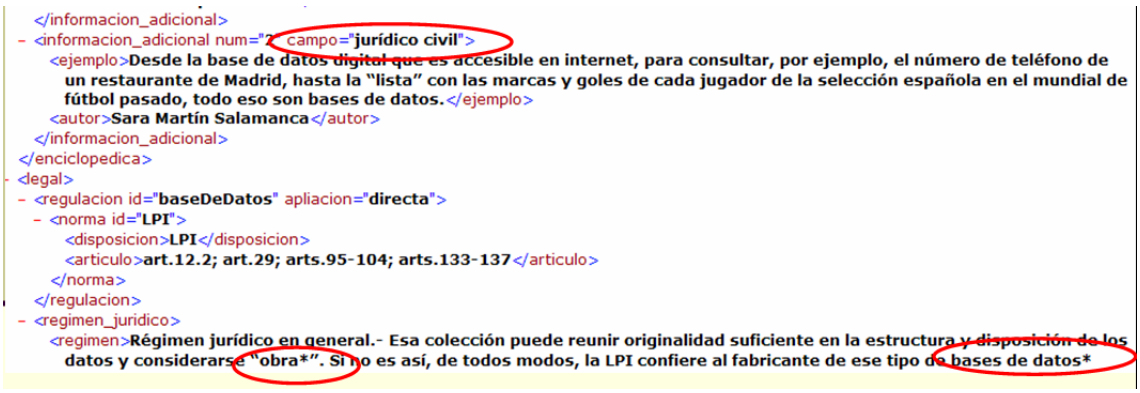
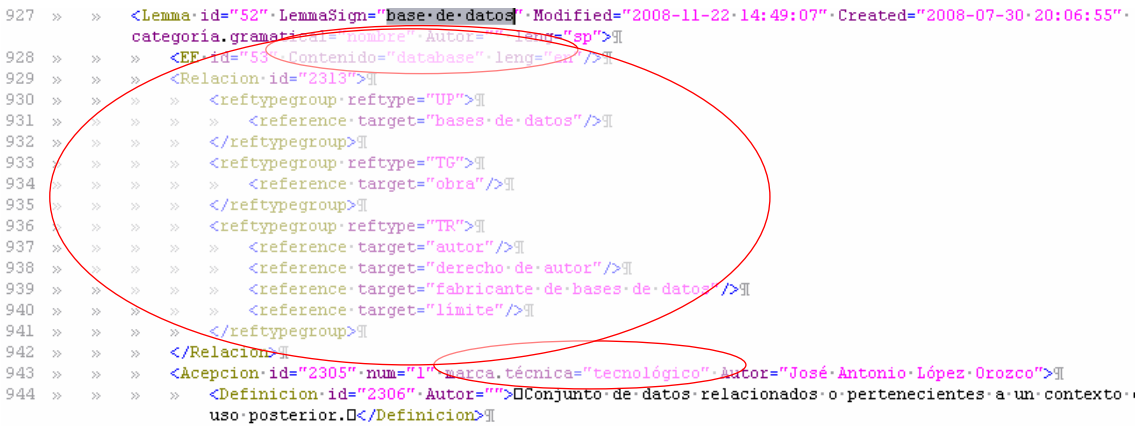


Figura 8.26. Términos relacionados con el lema de la entrada

Esta información se recogió en el glosario e-derecho en los elementos estructurales Lemma (en el atributo LemmaSign), Relación y Acepción (en el atributo marca.técnica) (figura 8.27).



Conjunto·de·datos·relacionados·o·pertenecientes·a·un·contexto·
uso·posterior.</Definicion>

Figura 8.27. Definición en la entrada (<Lemma>) de las relaciones semánticas del término lema ("base de datos")

Por lo tanto, cada entrada del glosario, incluía explícitamente una información semántica que es útil para la explotación del glosario: los términos organizados en categorías y relacionados semánticamente. La dimensión "Tesauro" del glosario mostraba esta información, pero de forma fragmentada y, en el caso de las relaciones semánticas, con redundancia:

1. sólo se puede acceder a las relaciones directas de cada término con otros. No existe una presentación sistemática, global del tesauro y esto hace difícil su explotación académica; y
2. los términos que participan en las relaciones se repiten en las entradas de los términos que tiene de destino. Por ejemplo, el término "base de datos", que es el lema de la entrada que se muestra en las figuras 8.26 y 8.27, se relaciona con 6 términos: bases de datos, obra, autor, derecho de autor, fabricante de bases de datos y límite. Esto significa que se repite una referencia con "base de datos" en cada una de estas 6 entradas.

Para poder obtener el máximo provecho de este tesauro fragmentado era necesario extraer de cada una de las entradas la información semántica de los elementos, Lemma, Acepción y Relación, y componerlo, como en un rompecabezas, de forma coherente para construir un sistema léxico independiente.

8.3.3. La construcción del tesauro e-derecho

La construcción se llevó a cabo aplicando el método HL, porque:

(1) se disponía de estructuras-t que podían ser utilizadas directamente para crear el tesauro;

(2) la composición de las estructuras-t se podía realizar de forma coherente gracias al modelo HL; y

(3) el proceso de construcción se podía realizar de forma inductiva desde las estructuras-t hasta el tesauro.

En consecuencia, para construir el tesauro académico e-derecho hemos aplicado la metodología HL de la forma siguiente:

Fase 1: identificación y definición del tipo de estructuras-t

Las estructuras-t se localizan dentro de las entradas del glosario que están marcadas con la etiqueta <Lemma>, en los elementos y atributos siguientes (figura 8.29):

1) el atributo Lemma.Sign del elemento Lemma;
2) el atributo marca.técnica de cada uno de los elementos Acepción; y
3) el elemento estructural *Relación*.

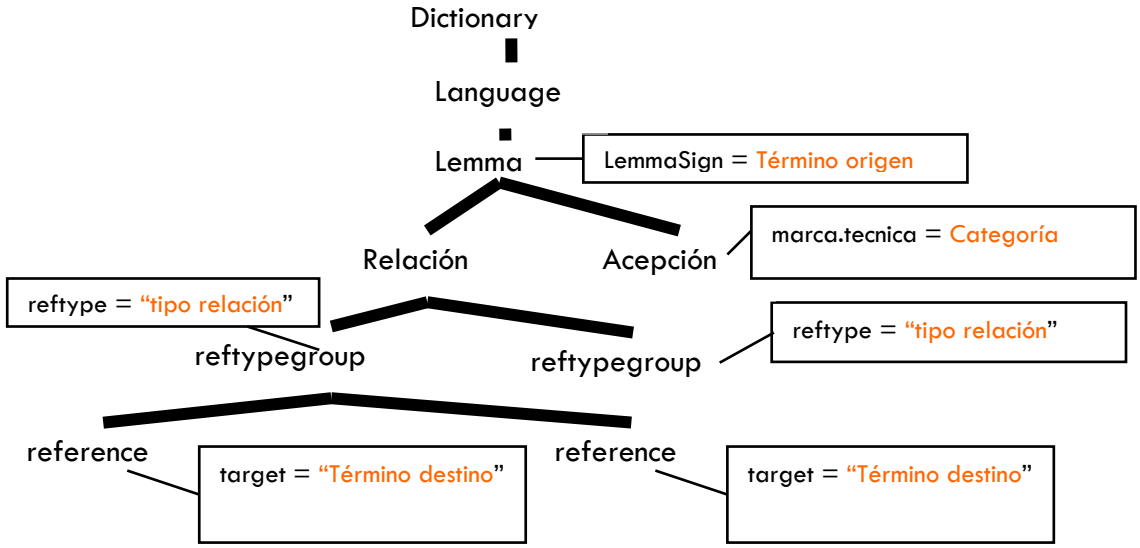


Figura 8.29 Definición gráfica de la estructura-t en el glosario e-derecho

Este lenguaje de marcas se define con la gramática siguiente (tabla 8.8):

Regla de producción	**Atributos**
Lemma → Relación, EE*, Acepción *, Regulacion*,Regimen_Juridico*,Jurisprudencia*	LemmaSign es el término origen de las relaciones
Acepción → Definicion?, Explicacion*	marca.tecnica es la categoría a la que pertenece el término del lema
Relación → reftypegroup+	id es un identificador único en todo el glosario para este grupo de relaciones

reftypegroup → reference +	reftypegroup.reftype es el tipo de relación semántica. En el glosario están definidas UP, USE, TG y TR
referente → λ	target es el término destino de la relación

Tabla 8.8. Gramática del lenguaje de marcas de las estructuras-t en el glosario e-derecho

1. el valor del atributo LemmaSign del elemento Lemma es el lema de la entrada, y es el término respecto del cual se especifican el resto de las relaciones semánticas que contiene el elemento Relación;
2. el valor del atributo marca.técnica del elemento Acepción es el ámbito temático en el que se define el significado. Se corresponde con la categoría a la que pertenece el término lema. Cuando una entrada tiene más de una acepción, es decir, es polisémica, y cada una de ellas está referida a un ámbito temático diferente, se recoge en el tesauro indicando la pertenencia del término lema a las diferentes categorías. De esa forma se desambigua el significado del término (figura 8.30); y

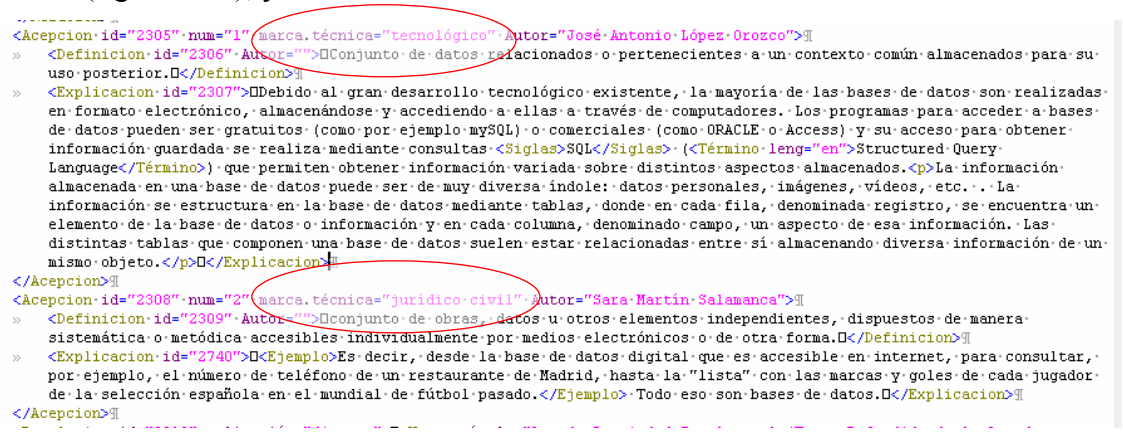


Figura 8.30. Polisemia del término base de datos

3. el elemento Relación contiene las relaciones semánticas del término lema con el resto de los términos. Se organiza en uno o varios *grupos de relaciones*, denominados *reftypegroup*. Cada grupo de relación se corresponde con un tipo de relación semántica, que se especifica en el atributo *reftype*. En la figuras 8.27 y 8.31 se pueden ver tres grupos de relaciones uno para el tipo "UP", otro para "TG" y el tercero para "TR". Cada grupo de relaciones contiene una referencia, *reference*, al término destino de la relación.

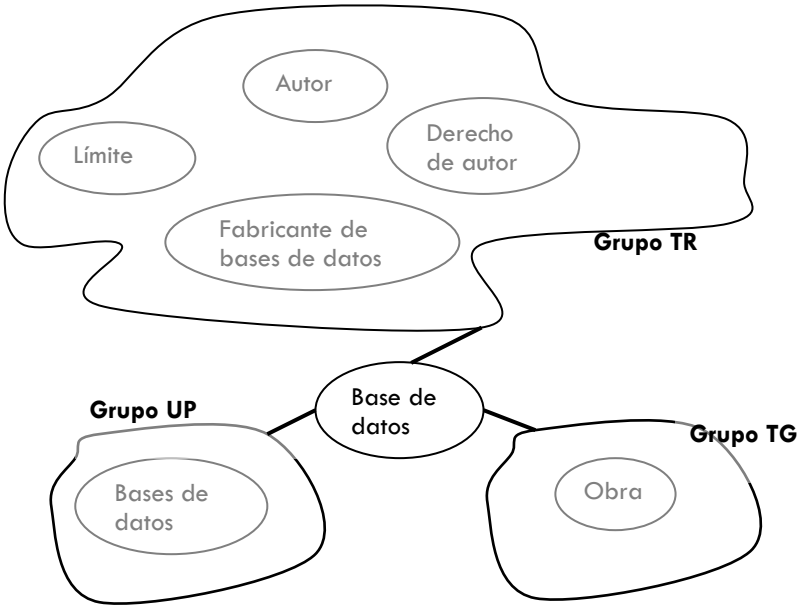


Figura 8.31. Relaciones semánticas del lema "base de datos"

Toda esta información está marcada con el lenguaje XML del modelo del glosario, por lo que es fácil de localizar y extraer. En la tabla 8.9 se muestran las etiquetas del glosario e-derecho que marcan las estructuras-t.

MARCA/PATRÓN		**COMPONENTE HL**
Etiqueta xml	**Atributo**	
Lemma		Estructura-t
Acepción	marca.técnica	CAT
Lemma	LemmaSign	TERM
reference	target	TERM*
Reftypegroup	reftype	REL[34]

Tabla 8.9. Definición del tipo de estructura-t en el glosario e-Derecho

Fase 2: extracción de las estructuras-t

Una vez definido el tipo de las estructuras-t y su localización, la siguiente operación extrae las estructuras-t del archivo XML.

Fase 3: análisis e interpretación de las estructuras-t

Se realiza aplicando la versión del algoritmo general de análisis HL que obtiene directamente las inserciones en el esquema de implementación de datos relacional HL (esquemas *hl_macro y hl_micro*[35]):

[34] REL es una variable que recoge el tipo de relación.
[35] Ver algoritmos en la sección anterior 8.2.3, quinta fase.

Entrada: texto etiquetado xml del glosario

Obtener las series semánticas;

Procesar Series de inclusión de CAT, si existen;

Procesar la inclusión de los TERM en CAT, si existen;

Procesar Series de TERM, si existen;

Salida: inserciones SQL

Aplicado, por ejemplo a la entrada de la figura 8.27, se obtendrían, en el primer paso, las series semánticas:

CAT= TECNOLÓGICO
TERM = base de datos
TERM= bases de datos
REL= UP
TERM =base de datos
TERM =obra
REL=TG
TERM =base de datos
TERM =autor
TERM =derecho de autor
TERM =fabricante de bases de datos
TERM =límite
REL=TR

La siguiente etapa del algoritmo generará las inserciones SQL. La tabla 8.10 muestra la traza del cálculo utilizando las series anteriores:

SERIES SEMÁNTICAS	**Procedimiento**	**SENTENCIAS SQL**
CAT= TECNOLÓGICO TERM = base de datos TERM= bases de datos REL= UP TERM =base de datos TERM =obra REL=TG TERM =base de datos TERM =autor TERM =derecho de autor TERM =fabricante de bases de datos TERM =límite REL=TR	Procesar CAT	insert into hl_micro (signo, tipo) VALUES (‘TECNOLÓGICO’, categoría);
	Procesar inserción de TERM en CAT	insert into hl_micro (signo, tipo) VALUES (’base de datos’, término); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’TECNOLÓGICO’,’base de datos’);
		insert into hl_micro (signo, tipo) VALUES (’bases de datos’, término); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’TECNOLÓGICO’,’bases de datos’);
		insert into hl_micro (signo, tipo) VALUES (’obra’, término); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’TECNOLÓGICO’,’obra’);
		insert into hl_micro (signo, tipo) VALUES (’autor’, término); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’TECNOLÓGICO’,’autor’);
		insert into hl_micro (signo, tipo) VALUES (’derecho de autor’, término); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’TECNOLÓGICO’,’derecho de autor’);

		insert into hl_micro (signo, tipo) VALUES (’fabricante de base de datos’, término); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’TECNOLÓGICO’,’fabricante de base de datos’);
		insert into hl_micro (signo, tipo) VALUES (’límite’, término); insert into hl_macro (rel ,signo1 ,signo2) values ('I',’TECNOLÓGICO’,’límite’);
	Procesar Series TERM	insert into hl_macro (rel ,signo1 ,signo2) values ('UP','base de datos’,’bases de datos’); insert into hl_macro (rel ,signo1 ,signo2) values ('TG','base de datos’,’obra’); insert into hl_macro (rel ,signo1 ,signo2) values ('TR','base de datos’,’autor’); insert into hl_macro (rel ,signo1 ,signo2) values ('TR','base de datos’,’derecho de autor’); insert into hl_macro (rel ,signo1 ,signo2) values ('TR','base de datos’,’fabricante de bases de datos’); insert into hl_macro (rel ,signo1 ,signo2) values ('TR','base de datos’,’límite’);

Tabla 8.10. Traza del algoritmo de inserción[36]

Fase cuarta: revisión y adecuación

Esta fase tiene como objetivo que el usuario pueda cambiar las tuplas obtenidas en la fase anterior e, incluso, que pueda añadir nuevas relaciones que considere, por su experiencia, que son interesantes para mejorar la eficacia del tesauro, por ejemplo establecer nuevas relaciones UP con los términos no preferidos. Los términos y las relaciones han sido revisados por los especialistas en Lexicografía[37]. Los cambios que se hicieron fueron, básicamente, (i) corrección de errores, por ejemplo, en los resultados del ejemplo (tabla 8.10) aparece un error en la relación (TG, ‘base de datos’, ‘obra’) que se corrigió por (TG, ‘obra’,’base de datos’); (ii) cambios de relaciones semánticas, por ejemplo, TG por TR; y (iii) cambios en las formas preferidas de los términos. Aunque esta fase permite mejorar el glosario, también puede suponer modificar el glosario para que sea coherente con el tesauro; por ejemplo cuando cambiamos las formas preferidas es necesario modificar en el glosario XML el valor del atributo LemmaSign por la nueva forma preferida.

[36] No considera la actualización de la frecuencia ni la asignación de un identificador, porque se realiza automáticamente, dependiendo de los valores ya almacenados en la base de datos.
[37] Miembros del Proyecto PIMCD 66/2008.

Fase quinta: inserción de las tuplas SQL

Consiste en ejecutar las sentencias obtenidas en la fase tercera, para incluir el nuevo conocimiento léxico en el HL del tesauro; por ejemplo, los resultados de la tabla 8.10 añadirán al esquema relacional HL las siguientes tuplas[38]:

En hl_micro: {(base de datos, término), (tecnológico, categoría), (jurídico civil, categoría), (bases de datos, término), (obra, término), (autor, término), (derecho de autor, término), (fabricante de base de datos, término), (límite, término)}

En hl_macro: {(I, tecnológico, base de datos), (I, jurídico civil, base de datos), (UP, base de datos, bases de datos), (TG, obra, base de datos), (TR, base de datos, autor), (TR, base de datos, derecho de autor), (TR, base de datos, fabricante de base de datos), (TR, base de datos, límite)}

Lo que significa que se añadirá al HL del tesauro *e-Derecho* las redes semánticas de la estructura-t del ejemplo (figura 8.32).

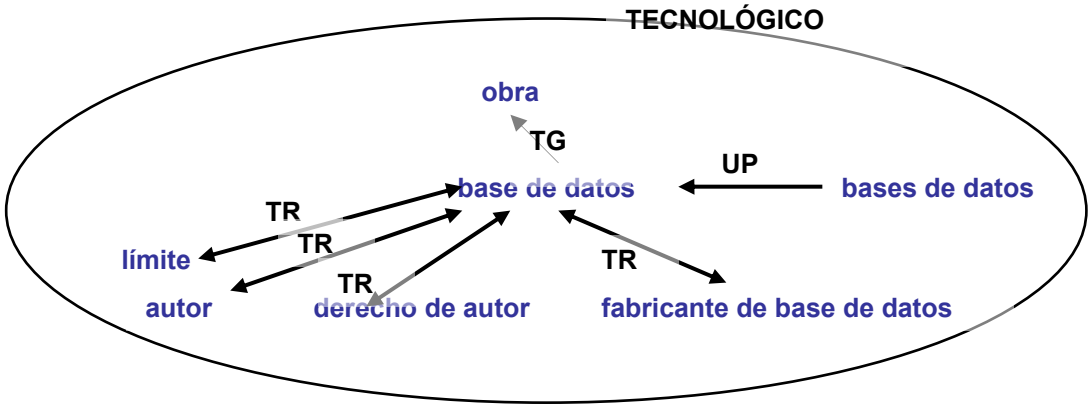


Figura 8.32. Interpretación gráfica HL resultado de la estructura-t de la figura 8.27

8.3.4. Resultados

El resultado es un tesauro con 369 términos organizados en 5 categorías principales y relacionados semánticamente con relaciones de equivalencia, USE, de hiperonimia, TG y de asociatividad, TR. El HL del tesauro e-derecho está construido en un SGBD[39] con el esquema de datos relacional descrito en este caso.

El esquema conceptual del tesauro es un HL con la categoría principal "tesauro", que contiene tres categorías únicas que son los campos respecto de los cuales se definen los

[38] Estamos considerando que, cuando se ejecutan estas sentencias de inserción, el sistema gestor asigna a cada tupla un identificador único y que actualiza las frecuencias de uso.

[39] Se ha utilizado para construir el prototipo el SGBD Oracle 10g Express Edition. Disponible en: http://www.oracle.com/lang/es/database/Express_Edition.html

términos: ‘jurídico civil’, ‘jurídico penal’ y ‘tecnológico’ (figura 8.33). Cuando el término no está definido o no se especifica el campo (marca técnica) de la definición el término sólo está incluido en la categoría principal “tesauro”.

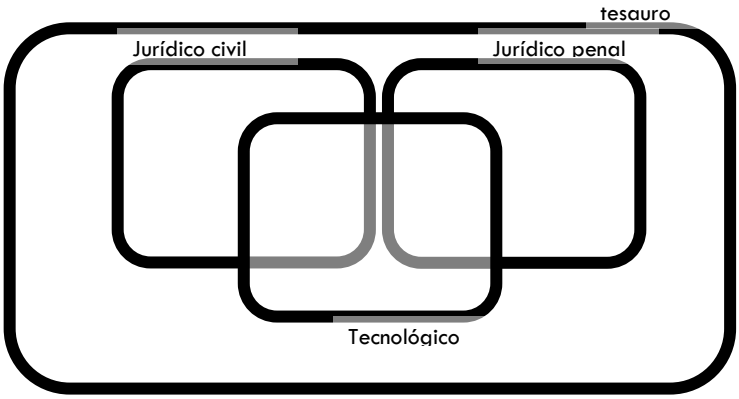


Figura 8.33. Macroestructura de categorías del tesauro e-Derecho

Los términos pueden ser términos preferidos y no preferidos. Los términos preferidos son los que tienen una entrada en el glosario. Los términos no preferidos sólo tienen en el glosario una remisión a los términos preferidos, por lo tanto están sin asignar**.** De los 369 términos, 348 son preferidos, tienen entrada en el glosario, y 21 no preferidos. Los términos no preferidos no están clasificados en ninguna categoría, excepto en la principal “tesauro”.

La relación semántica más utilizada es TR (153 relaciones) y organiza los términos en 38 redes de “familias semánticas de términos”. Existen 21 relaciones USE: 21 términos no preferidos con 17 términos preferidos (figura 8.34), y 6 jerarquías TG de sólo un nivel de profundidad con 9 relaciones.

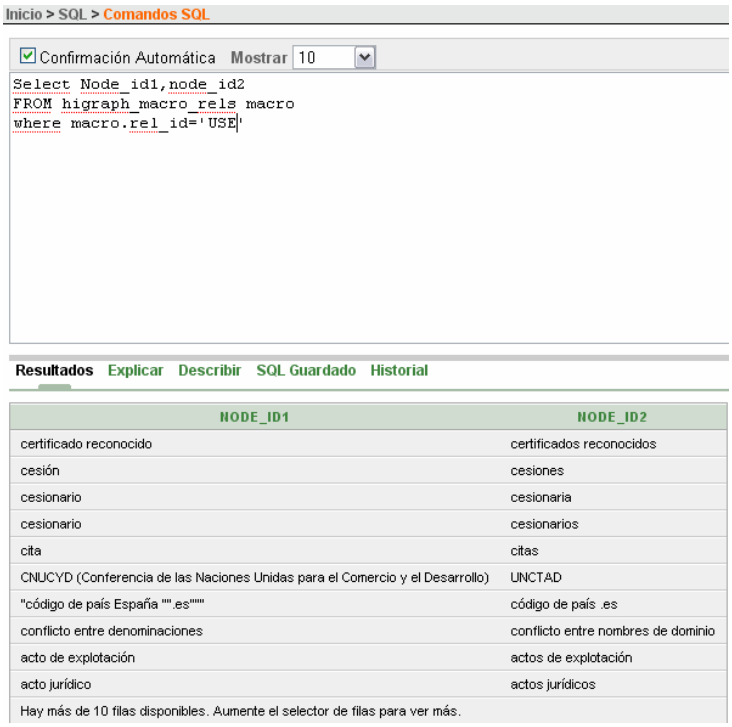


Figura 8.34. Consulta "relaciones semánticas USE" en el SGBD

8.3.5. Discusión

El tesauro del *glosario e-derecho* aporta un marco conceptual coherente para una terminología nueva y multidisciplinar que procede del campo del Derecho y de las Tecnologías de la Información y las Comunicaciones. Constituye un mecanismo eficaz para 1) entender este ámbito nuevo y todavía poco sistematizado; y 2) encontrar y/o ampliar la información en el glosario, que, de otra forma sólo dependería de la competencia del lector en la materia e-derecho, y en el uso de las herramientas de búsqueda exacta de caracteres, palabras o elementos estructurales.

El contenido del tesauro procede del glosario que, a su vez, ha sido creado a partir de una obra colaborativa realizada por un grupo de profesores y técnicos especialistas del derecho y de informática. En la elaboración de la obra original no se siguió un diseño o planta para estructurar la información, sino que fue escrita de forma libre a partir de términos seleccionados por los propios autores como significativos del ámbito. El glosario sistematiza este contenido multidisciplinar y poco estructurado, pero presenta una gran redundancia en el caso de las relaciones entre los términos, y carece de un esquema global de organización de los términos que no sea el alfabético y que permita al lector no experto acceder a la información que necesita. Esto resta eficacia al glosario, especialmente en su uso didáctico, y motiva la construcción de un tesauro específico para el glosario.

La metodología HL proporciona un procedimiento para la construcción inductiva, automática e incremental del tesauro con las estructuras-t que habían sido creadas por los profesores en el material original del glosario. En este caso práctico, la metodología se ha aplicado al contenido marcado del *glosario e-derecho*, del que pueden extraerse y reutilizarse las estructuras-t para crear el tesauro. El modelo HL proporciona el esquema conceptual y de implementación de datos general que se necesita para integrar las estructuras-t multidisciplinares en un sistema formal único, que es el HL. En consecuencia, para construir el tesauro no ha sido necesario llevar a cabo las fases de análisis, diseño y clasificación de los términos de las metodologías generales de construcción de tesauros.

Existen otras aproximaciones de construcción de tesauros a partir de contenidos que están basadas en extraer los términos y relaciones semánticas mediante técnicas de PLN (Paynter, et.al, 2000). Estas aproximaciones pueden utilizarse para extraer términos y relaciones semánticas de los textos, pero son más complejas y costosas de aplicar que el método HL. La metodología HL, por su parte, tiene un ámbito de aplicación más restringido, sólo a los contenidos con estructuras-t, y requiere de la intervención del usuario para definir el tipo de estructura-t; sin embargo, es más sencilla de aplicar y requiere un esfuerzo de programación menor.

Finalmente, este caso práctico demuestra que el modelo general HL puede utilizarse para integrar y gestionar uniformemente el conocimiento terminológico de áreas diferentes, con el apoyo de un SGBD para consultar, de forma flexible, el contenido del tesauro (figura 8.34), revisar, detectar y corregir errores y actualizar el contenido del tesauro.

8. 4. Resumen y conclusiones del capítulo

El objetivo de este capítulo de experimentación era comprobar la viabilidad, ventajas e inconvenientes del modelo y del método HL para construir tesauros académicos de explotación. Hemos aplicado el método HL a tres casos prácticos, (1) la especialización de tesauros generales a colecciones de materiales y recursos didácticos o de investigación; (2) la reconstrucción, como tesauro, de un índice temático de un repositorio de objetos virtuales académicos; y (3) la construcción de un tesauro académico de explotación para un glosario en línea. De los resultados de estas aplicaciones hemos extraído las siguientes conclusiones respecto del modelo HL: en

primer lugar, el modelo HL es viable, se puede aplicar a la construcción de los tesauros académicos de explotación de colecciones de recursos didácticos digitalizados, como por ejemplo los casos del tesauro ETB o CHASQUI, o de explotación de contenidos educativos, como el caso del tesauro e-derecho; en segundo lugar, el modelo HL es visual, lo que significa que los diagramas HL pueden ser utilizados como mapas terminológicos-conceptuales del ámbito del tesauro para (i) guiar al usuario en la búsqueda de los términos y para (ii) para ayudarle a entender y aprender los conceptos y términos de la especialidad que recoge el tesauro. En este sentido, hemos observado que los diagramas HL sirven para poner de manifiesto no sólo las estructuras semánticas explícitamente escritas en las estructuras-t, sino también aquellas que estaban implícitas en la colección global de estructuras-t que generan el tesauro y que eran desconocidas, incluso, para los propios autores de dichas colecciones; como hemos visto, por ejemplo, las redes de términos relacionados que sirven para definir los objetos virtuales del repositorio CHASQUI no estaban previamente definidas por el usuario; en tercer lugar, la generalidad del modelo HL permite un tratamiento uniforme del contenido del tesauro; finalmente, este tratamiento uniforme puede ser aprovechado para construir herramientas software de carácter general que permitan la construcción, actualización y gestión de los tesauros, ventaja que no hemos podido demostrar en este trabajo de tesis porque la magnitud de la tarea sobrepasa las posibilidades de una sola persona o de un grupo reducido de personas.

Respecto de la metodología HL, hemos comprobado su viabilidad para la construcción inductiva e incremental de los tesauros académicos de explotación en los tres casos prácticos. Hemos comprobado que, utilizando las estructuras-t escritas por los profesores en sus materiales didácticos y esta nueva metodología HL, se pueden construir los tesauros académicos de explotación que describen de forma precisa estos materiales o el contenido de estos materiales. Sin embargo, también hemos comprobado que la intervención del profesor en la primera fase del método, la identificación de las estructuras-t, es clave porque resuelve el problema de la diversidad de estructuras-t que aparece, incluso, si nos restringimos a un mismo tipo de estructura-t estándar; por ejemplo el caso 1 y 2 demuestra cómo el estándar IEEE-LOM es interpretado y utilizado de forma diferente por los equipos docentes. Además, también hemos podido comprobar que el método HL simplifica el proceso de construcción, siempre que sea aplicable, porque no es necesario abordar las fases de (i) búsqueda y selección de fuentes de términos, (ii) el análisis de los términos extraídos de las fuentes y (iii) el

diseño del esquema conceptual del tesauro. Como también hemos comprobado, el método HL realiza, de forma iterativa, la extracción de nuevas estructura-t de los metadatos, el análisis de los componentes de las estructuras-t, y la inserción de estos componentes en el tesauro, de forma que el tesauro va creciendo con los términos y relaciones semánticas obtenidos de las estructuras-t de forma incremental y continuada. En este proceso no se distingue entre las etapas de construcción y de mantenimiento, así que, desde el punto de vista del método HL, el tesauro está en continua construcción.

Finalmente, hemos observado que la calidad de los tesauros construidos con el método HL depende de la calidad y cantidad de estructuras-t, pero, en cualquier caso, la aplicación del método garantiza: (i) la familiaridad del usuario, que son los docentes y estudiantes, con la terminología del tesauro, y (ii) la precisión para describir los objetos, contenidos o recursos, de los que provienen las estructuras-t.

Capítulo 9

Recapitulación, conclusiones finales y líneas trabajo futuro

"No se trata simplemente de hallar la respuesta correcta, sino más bien de comprender porqué existe una respuesta, si la hay, y por qué dicha respuesta presenta una determinada forma" (Stewart, 2004)

En este último capítulo presentamos la recapitulación de toda la investigación que hemos realizado y, a partir de ella, elaboramos las conclusiones finales y examinamos cuáles son las líneas de trabajo que nos hemos ido dejando en el camino por no haber podido abordarlas o por tener un contenido que no estaba directamente relacionado con los objetivos de la tesis, pero que pueden servir para abrir nuevas líneas de investigación.

9.1 Recapitulación

En este apartado retrocedemos al inicio de este trabajo de investigación para volver a recorrer, de nuevo, el camino ya andado, y así, como propone Stewart (2004), comprender, con más perspectiva, por qué existe la respuesta que proponemos y por qué esta respuesta presenta la forma que hemos desarrollado.

9.1.1. Objeto de estudio

El objeto de esta investigación son los tesauros académicos de explotación, un nuevo tipo de tesauros que está surgiendo en los entornos electrónicos de formación universitaria, *e-learning,* para expresar y organizar el conocimiento y las creaciones intelectuales de los actores del mundo académico: profesores, investigadores y estudiantes. Los tesauros académicos de explotación son el instrumento lingüístico para sistematizar, expresar con un lenguaje cercano, y manipular el conocimiento contenido en los materiales y recursos didácticos y de investigación digitales, desarrollado o recopilado por los equipos docentes. Este tipo de tesauros es útil, fundamentalmente, para (i) ayudar al profesor e investigador a organizar conceptualmente sus materiales, haciendo más fácil la localización, selección y uso; y (ii) ayudar al estudiante a entender

y aprender los conceptos, y a usar de forma adecuada la lengua de especialidad de la disciplina o área de conocimiento que cubra el tesauro.

9.1.2. Cuestiones de investigación

Las cuestiones que se plantean en este trabajo de investigación son dos:

(1) cómo definir esta nueva forma de tesauro que, a diferencia de los tesauros de referencia creados y utilizados en Informática, Biblioteconomía y Documentación, y Lingüística, se ajusta al dominio de conocimiento, al lenguaje y a las necesidades concretas del usuario. El usuario es el profesor, investigador o estudiante; el dominio de conocimiento corresponde a las disciplinas o áreas de conocimiento inter e intradisciplinares sobre las que se investiga, se enseña y se aprende; y el lenguaje es el de especialidad que el equipo docente e investigador utiliza para expresar el conocimiento del dominio en sus materiales didácticos y de investigación. Con este lenguaje construye pequeñas redes de términos relacionados semánticamente que sirven para clasificar y describir los objetos del ámbito del tesauro, que son los materiales o recursos que se utilizan para la docencia e investigación. Llamamos a estas estructuras de términos en semántica libre, estructuras-t, porque son porciones de la lengua general, o de especialidad, de los profesores y son creadas de forma libre, no consolidada, para relacionar un material o un recurso con el conocimiento compartido general sobre una disciplina o área. Constituyen una forma simplificada de proporcionar semántica, explicar y/o clasificar, el contenido de estos recursos en las aplicaciones informáticas que los gestionan como, por ejemplo, las bibliotecas digitales, las bases de datos documentales o los repositorios de objetos de aprendizaje.

(2) La segunda cuestión es cómo definir los mecanismos que permitan a los usuarios construir y mantener actualizados estos tesauros de la forma más sencilla posible. Cuando hablamos de mecanismos nos referimos a: (i) cómo representar de forma general el conocimiento léxico de un tesauro de cualquier ámbito; y (ii) cómo construirlos y mantenerlos de forma sistemática.

Responder a estas cuestiones significa, en cuanto a la primera pregunta (1) encontrar una definición, y ejemplos, de tesauros académicos de explotación; en cuanto a la segunda cuestión (2) se trata, en primer lugar, de encontrar un meta-modelo, o modelo general, que sistematice uniformemente el contenido de estos tesauros, y, en segundo

lugar, encontrar un método inductivo que construya y actualice el tesauro inductivamente recogiendo y organizando, con el modelo general, las estructuras-t de los materiales o recursos que constituyen el ámbito del tesauro. Éstos mecanismos, el modelo y método, son el fundamento para construir aplicaciones informáticas generales que faciliten, a los equipos docentes, la construcción y actualización de sus tesauros académicos de explotación.

9.1.3. Hipótesis de trabajo

La respuesta que hemos dado en esta investigación está basada en la hipótesis de considerar que se dan las cuatro características siguientes: (i) el tesauro es un sistema estructurado de signos lingüísticos en el que el valor del significado de cada signo depende de su posición diferencial respecto de los demás; (ii) el tesauro representa las nociones de un dominio de conocimiento mediante signos organizados en grupos según los tipos de relaciones semánticas de generalización-especialización, equivalencia, asociación y, posiblemente, otras relaciones específicas del dominio; (iii) existe un modelo formal único para representar esta concepción de tesauro; y (iv) existen estructuras de términos en semántica libre que representan parcialmente, las nociones de un dominio mediante signos organizados en grupos por las relaciones semánticas mencionadas y que son creadas libremente por los profesores para describir y clasificar sus creaciones intelectuales. Si se dan estas cuatro características, entonces se logran los dos objetivos siguientes (i) es posible sistematizar, con el modelo anterior, de forma general y uniforme el contenido de cualquier tesauro y los procesos de construcción y gestión del tesauro; y en consecuencia, (ii) es posible definir el mecanismo para construir y actualizar de forma sistemática los tesauros académicos de explotación de forma inductiva a partir de las estructuras-t creadas por los docentes para describir y clasificar los materiales, contenidos y recursos didácticos en los entornos *e-learning*.

9.1.4. Análisis crítico del estado de la cuestión

A partir de esta hipótesis hemos realizado un análisis para: (1) establecer la naturaleza y aplicaciones de los tesauros de explotación que preceden a los nuevos tesauros académicos de explotación en el contexto del *e-learnig*; (2) establecer el contexto actual de trabajo académico *e-learning* donde emergen los tesauros académicos de explotación; (3) definir cómo son y dónde se localizan algunas de las estructuras-t que sirven de fuente a los tesauros académicos de explotación; (4) la evaluación de los

modelos utilizados para construir tesauros; y (5) definir los métodos de construcción de los tesauros de explotación.

9.1.4.1 Naturaleza y aplicaciones de los tesauros de explotación

El primer punto de este análisis, establecer la naturaleza y aplicaciones de los tesauros de explotación, nos ha llevado a revisar el concepto más general e interdisciplinar de vocabulario. El vocabulario es, desde el punto de vista lingüístico, el conjunto de palabras de una lengua o de un dominio determinado, pero también la obra lexicográfica que las recoge. Desde el punto de vista de la Lingüística Computacional, el vocabulario es un componente de los sistemas de procesamiento del lenguaje natural que, además de palabras, contiene información sobre las relaciones entre ellas, sus propiedades, etc. En Biblioteconomía y Documentación y RI se entiende por vocabulario un lenguaje documental, es decir, un conjunto de palabras controlado por unas convenciones que determinan las formas sintácticas y los significados únicos de cada término para evitar la redundancia y la polisemia. Sirven como sistemas de referencia universales para clasificar las obras bibliográficas y documentales. La Tecnología Web y el *e-learning* consideran que los vocabularios son recursos lingüísticos que contienen términos y conceptos relacionados y cuyo fin es la representación semántica de los objetos digitales con contenido de la Web y de los entornos virtuales de aprendizaje, respectivamente.

Las relaciones semánticas definen el significado de los términos en los vocabularios orientados a la explotación porque permiten expandir o precisar las consultas, utilizando estas relaciones y los términos de las consultas. La Lingüística proporciona la definición y clasificación de los distintos tipos de relaciones semánticas, lo que es imprescindible para definir mecanismos de explotación uniformes. En semántica se consideran cuatro tipos de relaciones clásicas: (i) jerarquía, (ii) inclusión, (iii) equivalencia, (iv) oposición y otras, como las de asociación semántica. Los vocabularios orientados a la explotación utilizan, fundamentalmente, las relaciones de equivalencia, de inclusión, y asociativas. En la búsqueda de información o recursos, las relaciones semánticas proporcionan una estructura organizada de términos que guía al usuario en (i) su exploración del dominio de información o de la colección de objetos de contenido; (ii) la recuperación, permitiendo ampliar o restringir la expresión de búsqueda que introduce el usuario en su consulta; y (iii) el filtrado de la consulta, según las preferencias del usuario, que proporciona el conjunto de términos que están relacionados con su perfil.

La efectividad de un vocabulario de explotación se mide con dos parámetros: (i) la exhaustividad, o capacidad de recuperar la información u objetos más relevantes para el usuario en un proceso de consulta, y (ii) la precisión, o capacidad de obtener el mínimo número de objetos no relevantes en un proceso de consulta. La exhaustividad y la precisión son parámetros inversamente proporcionales. Los vocabularios muy específicos mejoran la precisión en las búsquedas, pero reducen la exhaustividad; por el contrario, los vocabularios con términos más generales benefician la exhaustividad pero empeora la precisión de los resultados de búsqueda. Otro factor que influye en la efectividad de los vocabularios de explotación es el control de la sinonimia y ambigüedad. Los sinónimos son imprescindibles para arreglar los desajustes que siempre existen entre el vocabulario utilizado en la indexación y el vocabulario de los usuarios que consultan pero, si no están bien relacionados, se reducen la exhaustividad y la precisión. La ambigüedad reduce siempre los resultados en precisión.

En *e-learning* los vocabularios se utilizan, principalmente, para la explotación didáctica de recursos digitalizados. Los recursos didácticos digitalizados se pueden clasificar con los términos de uno o varios vocabularios pero, también, con los metadatos, que son el conjunto de propiedades y valores que caracterizan a un recurso. Si se utilizan metadatos conjuntamente con vocabularios de referencia reconocidos, éstos proporcionan términos para expresar de forma regular y universal los valores de las propiedades de los metadatos. El problema, sin embargo, es que no siempre los vocabularios de referencia (i) representan con precisión las colecciones de recursos educativos, ni (ii) se ajustan al propósito de uso docente e investigador propios de la Universidad, y tampoco (iii) se ajustan al vocabulario utilizado por los docentes e investigadores. Para terminar este primer análisis hemos estudiado la clasificación de los vocabularios de explotación propuesta por el estándar de construcción de tesauros de explotación ANSI/NISO Z39.19 (2005), utilizando ejemplos reales de aplicación extraídos del contexto del *e-learning*. Así, hemos distinguido cinco tipos de vocabularios de explotación con aplicación al *e-learning*: (i) los vocabularios simples, que son listas cerradas de términos sin relaciones semánticas; (ii) las clasificaciones y taxonomías, que son vocabularios que contienen únicamente relaciones de tipo inclusión, (iii) los tesauros, vocabularios que contienen, básicamente, relaciones de tipo inclusión, equivalencia y asociación, e incluso pueden contener relaciones específicas del dominio de la información, (iv) las ontologías, que son conceptos relacionados que pueden tener definidos mecanismos de inferencia para razonar y que, cuando se asocian

con términos, se llaman ontologías léxicas; y (v) finalmente, los diccionarios y glosarios que son vocabularios orientados al uso humano y poco adecuados a la explotación automática porque la información que contienen sobre los términos está expresada en lenguaje natural.

9.1.4.2. Contexto de trabajo académico del *e-learning*

En el segundo análisis que hemos realizado, sobre el contexto de trabajo académico del *e-learning*, hemos dado respuesta a las preguntas: qué son, cómo son, cómo funcionan y qué aportan los actuales entornos virtuales de enseñanza y aprendizaje denominados campus virtuales. Estos entornos facilitan al profesor la creación y difusión del conocimiento y, probablemente por eso, en ellos está surgiendo la necesidad de nuevos tesauros académicos para la explotación de los recursos didácticos que contienen este conocimiento.

Los campus virtuales son espacios en Internet creados con aplicaciones Web, principalmente plataformas *e-learning*, con un propósito educativo. Existen otros términos, que hemos analizado, para denotar campus virtual, plataformas *e-learning* y enseñanza-aprendizaje virtual, aunque no siempre tienen el mismo significado. Las plataformas *e-learning* permiten crear lo que hemos denominado espacios de aprendizaje, que son aulas electrónicas donde trabajan los profesores con sus alumnos utilizando las herramientas que la plataforma pone a su disposición y que, de forma genérica, les permiten: (i) administrar sus usuarios, por ejemplo, los estudiantes y profesores ayudantes, (ii) comunicarse, (iii) gestionar y publicar los recursos didácticos, (iv) trabajar en grupo y (v) evaluar a los estudiantes. Otras plataformas más específicas tienen funcionalidades añadidas como, por ejemplo, simuladores, itinerarios de aprendizaje, interacción síncrona, etc.

En realidad, puede considerarse que un campus virtual es el conjunto de los espacios de aprendizaje de todos sus profesores. Dependiendo de la estrategia elegida para crear el campus virtual, tecnológica, institucional o didáctica, éste tendrá una forma más estática y regular o más dinámica e irregular. Cuando en los campus virtuales se priman la tecnología o los objetivos institucionales frente a los objetivos didácticos, el profesor tiene poca participación en la creación de sus espacios de aprendizaje y, en consecuencia, el campus virtual es más rígido, limitado y regular en su forma y funcionamiento, pero más sencillo de crear y mantener. Sin embargo, cuando el objetivo es potenciar al máximo las posibilidades didácticas, el profesor tiene un papel muy

activo, ya que es el creador y responsable de sus espacios de aprendizaje, y el campus virtual es dinámico, abierto, rico en funcionalidades y formas. En cualquiera de los casos, el campus virtual se construye con una o varias plataformas *e-learning* de gestión de espacios de aprendizaje, conocidas como LMS y LCMS. Además, puede tener integrados LMS específicos, herramientas satélites, portales Web educativos, y repositorios educativos. Los repositorios educativos son contenedores de recursos educativos en los que se almacena y accede a dichos recursos mediante metadatos y vocabularios.

El modo en el que los profesores utilizan los campus virtuales depende fundamentalmente de su experiencia. Considerando un campus virtual centrado en el profesor, hemos distinguido tres tipos de usos didácticos: (i) en las etapas iniciales, el uso básico de los espacios de aprendizaje como páginas Web para la difusión del conocimiento; (ii) en una etapa intermedia, el uso como extensión de las clases presenciales, en el que no sólo se difunde el conocimiento, sino que también se comparte en grupos de trabajos, se discute en los espacios de comunicación, y se evalúa con las herramientas de evaluación; finalmente, hemos distinguido (iii) una etapa de uso avanzado, cuando el profesor con experiencia empieza a aplicar didácticas diferentes de las utilizadas en las clases presenciales, aprovechando, de verdad, el potencial del *e-learning* para personalizar el aprendizaje, favorecer los modelos socioconstructivistas y autónomos de aprendizaje, formas, todos ellos, que preparan al estudiante para continuar formándose a lo largo de la vida. En las etapas de experiencia media y avanzada, los repositorios educativos tienen un papel clave para la creación, difusión y compartición del conocimiento, pero su uso es, actualmente, muy limitado debido a la dificultad que implica, para los profesores y estudiantes, describir con metadatos y vocabularios sus recursos educativos. Supone un esfuerzo considerable buscar y conocer en profundidad los vocabularios de referencia para utilizarlos correctamente y, con frecuencia, no es rentable porque estos vocabularios no están preparados para describir los contenidos de los recursos educativos con la precisión y la forma que los profesores consideran adecuada para sus propósitos.

9.1.4.3. Estructuras-t

La tercera parte del análisis nos ha llevado a definir cómo son y dónde se localizan las estructuras-t que pueden servir de fuente a la construcción de los tesauros académicos de explotación. Partiendo de los resultados de los análisis anteriores, que han servido

para definir el papel de los tesauros de explotación en el contexto actual del *e-learning* como sistemas lingüísticos para la representación conceptual de los contenidos de estos recursos, hemos localizado una de las principales fuentes de estructuras-t que son las propiedades de clasificación temática de los metadatos de los recursos didácticos. En concreto, en la recomendación de los metadatos estándar LOM del *e-learning*, se aplican tesauros y taxonomías para dar valores a la propiedad de clasificación. La forma de uso, que está también estipulada en el estándar, consiste en seleccionar, de los tesauros de referencia, tantas secuencias de términos relacionados por inclusión, denominadas caminos taxonómicos, como se necesiten para clasificar correctamente el recurso. Por lo tanto, una posible fuente de estructuras-t son los caminos taxonómicos que se localizan en la propiedad clasificación de los metadatos LOM de los recursos educativos. El problema es que para los profesores no es sencillo crear estas estructuras-t porque se necesita tener un conocimiento amplio sobre el contenido del recurso, los metadatos LOM y los vocabularios. Esto explica por qué, en vez de aplicar los vocabularios de referencia utilizan su propio lenguaje para crear las estructuras-t que documentan los recursos, de acuerdo con su concepción del dominio de conocimiento y de los contenidos de los recursos didácticos.

9.1.4.4. Modelos para la construcción de tesauros

El siguiente punto que hemos analizado son los modelos utilizados para construir tesauros. El objetivo es verificar si se cumple el tercer enunciado de nuestra hipótesis: existe un modelo formal único para representar formalmente el tesauro concebido como un sistema de signos estructurado por relaciones semánticas de equivalencia, inclusión y asociación, en el que el valor del significado de un signo depende de su posición diferencial respecto de los demás y cuyo contenido describe las nociones de un dominio de conocimiento. Hemos organizado el análisis en cuatro partes: la primera parte, es el análisis de las características y requisitos propios de los tesauros de explotación; la segunda parte, el modelo de los estándares de construcción de tesauros monolingües de explotación, que propone una forma común de sistematizar el contenido, la presentación y el mantenimiento de estos tesauros; la tercera parte, los modelos tradicionales de construcción de tesauros, alfabético y sistemático; y la cuarta parte, los modelos de datos informáticos utilizados para construir tesauros.

De la primera parte hemos extraído las características propias del contenido de los tesauros de explotación: (i) altamente relacionados, (ii) en permanente evolución, (iii)

formados por términos, relaciones semánticas y categorías, (iv) con cinco formas de presentación: alfabética, índice permutado, jerárquica, sistemática y gráfica; y (v) con una estructura modular: la microestructura, que contiene sólo la información relativa a un término, y la macroestructura, que es la estructura global del tesauro formada por todas las microestructuras. Además, los requisitos que deben cumplir los tesauros de explotación son los siguientes: (i) el control de la sinonimia y la ambigüedad; (ii) la utilización precisa de las relaciones y los términos que representan las nociones del dominio; (iii) la adecuación a la cobertura, el alcance, y el tipo de usuario que se hayan definido en el plan del tesauro; (iv) la capacidad de expresar de forma precisa y exhaustiva el dominio de conocimiento, (v) la capacidad de adaptarse a los cambios permanentes en el lenguaje, (vi) la efectividad para la explotación de información o de colecciones de recursos didácticos, y (vii) la accesibilidad.

La segunda parte del análisis nos ha permitido establecer el modelo estándar que define el contenido de los tesauros de explotación de forma universal. Este modelo se ha ido construyendo de forma empírica durante más de sesenta años[1] a partir de una diversidad de estándares procedentes de Estados Unidos, el estándar ANSI/NISO Z39.19, de la organización internacional UNESCO, el ISO 2788, de Gran Bretaña, el BS-5723, y de España, la UNE 50106, 1990 y 1995, que en realidad es una versión para el español del ISO 2788. Actualmente, todos los estándares definen el contenido del tesauro con: (i) términos, elegidos de acuerdo a unas reglas de estandarización; (ii) categorías, que asocian los términos con criterios semánticos o estadísticos en un mismo nivel jerárquico; (iii) relaciones semánticas que, básicamente, son de tipo inclusión, término genérico-específico TG-TE, de tipo equivalente, con mención del término preferido "use término" o "use término para", USE-USEPARA, y de asociación, término relacionado TR; (iv) objetos de contenido, que son los objetos clasificados y descritos con el tesauro; y (v) índices, que son estructuras que relacionan los términos del tesauro con los objetos de contenido. El modelo estándar también define los modos de presentación del contenido que han demostrado a lo largo de los años ser eficaces para acceder al contenido. Finalmente, el modelo define un conjunto de reglas para controlar que las modificaciones se realicen sin perder la consistencia del contenido. Sin embargo, las modificaciones que afectan mucho a la estructura no pueden llevarse a cabo más que rehaciendo el tesauro.

[1] Primero en papel y después en formato electrónico.

En definitiva, la sistematización estándar tiene como fin (i) garantizar que el tesauro sea un sistema eficaz de indexación y búsqueda de objetos de contenido, y (ii) establecer una nomenclatura común que permita el intercambio y reutilización del contenido de los tesauros y de los objetos de contenido. Debe constituir, por lo tanto, un referente para los modelos de datos informáticos que se utilicen en la construcción de los tesauros de explotación.

La tercera parte del análisis nos ha permitido definir cuáles son los modelos de construcción de los tesauros tradicionales, que reproducen los modos acostumbrados de presentación alfabética y sistemática en la organización interna del tesauro. El modelo alfabético organiza el contenido en una lista ordenada alfabéticamente de microestructuras de términos. Es el modelo más antiguo, el más sencillo, en general, de aplicar y el más barato, pero tiene tres inconvenientes importantes: (i) el usuario tiene que conocer exactamente los términos que necesita, (ii) muestra el contenido del tesauro fragmentado, y (iii) es difícil mantener la consistencia del contenido ya que el esquema global está implícito y porque la información de las relaciones tiene que repetirse en las entradas de cada uno de los términos que participan en la relación. Es un modelo adecuado para tesauros con macroestructuras sencillas, que tengan pocas actualizaciones, y que se van a presentar en soporte de papel o digital.

El modelo sistemático organiza el contenido con el esquema conceptual del dominio que, de forma deductiva o inductiva, se ha creado a partir de un análisis del dominio. Es un modelo más elaborado que el alfabético, que exige un esfuerzo mayor de diseño e implementación, y que no es fácil de actualizar cuando las modificaciones afectan al esquema. Sin embargo presenta tres ventajas: (i) aporta un esquema terminológico-conceptual del ámbito del tesauro que facilita la comprensión del dominio y la búsqueda, aplicando múltiples estrategias basadas en la postcoordinación o precoordinación de términos y en la navegación; (ii) facilita el control de la consistencia del contenido del tesauro, tanto en la construcción como en el mantenimiento del mismo; y (iii) facilita la interoperatividad, porque el esquema del tesauro permite interpretar e intercambiar el contenido entre diferentes centros o aplicaciones. Es un modelo adecuado para los tesauros de explotación siempre que se reproduzca el esquema sistemático en un esquema de datos informático y que no se vayan a realizar actualizaciones que afecten a este esquema.

Finalmente, la última parte de nuestro análisis sobre modelos de tesauros se ha centrado en los modelos de datos informáticos aplicados a la construcción de los tesauros. Es el

más largo en extensión porque cubre doce modelos: (i) redes semánticas, (ii) hipertexto, (iii) higraphs, (iv) entidad-relación, (v) entidad-relación extendido, (vi) orientado a objetos, (vii) relacional, (viii) RDF y RDFS, (ix) OWL, (x) SKOS-Core, (xi) IMS VDEX, y (xii) CEN XVD. En cada uno de ellos hemos analizado los fundamentos teóricos, algunas aplicaciones reales, y su adecuación como modelo general para la representación sistemática del contenido de cualquier tesauro.

Los modelos de datos se dividen en (i) modelos conceptuales, que sirven para describir, a nivel lógico, la organización del contenido, y (ii) modelos de implementación de datos, que representan la organización de los datos a nivel de las aplicaciones software. Esta distinción es importante porque, en informática, el diseño de sistemas se realiza en dos fases secuenciales: en la primera, que es la de diseño conceptual, se obtiene el esquema lógico que representa de forma abstracta todo el sistema, y en la segunda, la fase de implementación, se obtiene un esquema de implementación de datos que es el que manipulará la aplicación de gestión de datos. Los modelos conceptuales para el diseño de tesauros que hemos analizado son: las redes semánticas, el hipertexto, los higraph, el modelo Entidad-Relación, el Entidad-Relación Extendido, y el modelo orientado a objetos, que es el menos abstracto de todos porque utiliza las estructuras de datos de los lenguajes de programación orientados a objetos.

Las redes semánticas y el hipertexto utilizan el formalismo matemático de los grafos para representar el contenido del tesauro. En las redes semánticas los nodos contienen términos, categorías o valores de las propiedades de los términos, por ejemplo las notas de ámbito; los arcos etiquetados representan relaciones semánticas. En el hipertexto los nodos son más complejos y contienen toda la microestructura de los términos; los arcos son las relaciones semánticas.

El modelo higraph es un modelo matemático y visual que tiene un componente estructural y un componente semántico. El componente estructural es un modelo basado en grafos, pero en el que los nodos son conjuntos; esto implica que los nodos pueden contener otros nodos-conjuntos a los que se pueden aplicar las operaciones de conjuntos definidas en matemáticas. Los nodos que no contienen elementos se denominan nodos atómicos, y son los elementos básicos de la estructura. Los arcos pueden ser binarios, como en los grafos que conectan dos nodos, pero también múltiples cuando conectan muchos nodos todos con todos. Este último tipo de arco se denomina hiperarco. El segundo componente, el modelo semántico, define la semántica de un higraph mediante la composición incremental desde el significado de los nodos atómicos hasta el

significado global del higraph. El significado de los nodos atómicos se define mediante una función de interpretación que toma valores de un dominio de significados y se lo asigna a los nodos atómicos de forma que nunca dos nodos atómicos pueden compartir un significado. El significado de los nodos no atómicos se calcula combinando de forma incremental los significados de sus nodos componentes. El tipo de combinación depende de la forma en que se descomponen los nodos no atómicos en sus componentes: si es por descomposición ortogonal, se componen mediante el producto cartesiano no ordenado de los significados de sus componentes, y, si es por inclusión, se componen mediante al unión de los significados de los componentes.

En el modelo Entidad-Relación se utiliza, también, una notación gráfica para representar las entidades -con rectángulos-, los atributos -con elipses- y las relaciones -con rombos-; en el tesauro los términos y categorías son entidades, las notas de ámbito son propiedades y las relaciones semánticas son relaciones del sistema.

El modelo Entidad-Relación Extendido añade al modelo Entidad-Relación algunas posibilidades de representación más, como, por ejemplo, la representación de tipos, restricciones de participación y cardinalidad. Estos dos últimos modelos se utilizan, principalmente, para el diseño conceptual de bases de datos relacionales.

El modelo Orientado a Objetos se basa en las estructuras de los lenguajes de programación orientada a objetos para definir no sólo la estructura de los datos sino también las operaciones. Utiliza, entre otras, las nociones de objeto, que es una entidad; atributos, que son las propiedades de los objetos; métodos, que son las operaciones entre los objetos; y clases, que sirven para definir los tipos de objetos.

Todos los modelos conceptuales que hemos analizado, excepto el modelo higraph, tienen, desde el punto de vista de nuestra hipótesis de modelo único para la representación de los tesauros, dos limitaciones importantes (1) no existe la noción de conjunto para definir las categorías del tesauro, que son conjuntos de términos con las operaciones y propiedades propias de los conjuntos, y (2) no existe la noción de hiperarco para definir las relaciones de asociación TR entre varios términos que están relacionados todos con todos. Aunque existen modos de resolver estas limitaciones, los esquemas resultantes (i) no son generales, dependen del ámbito de cada tesauro y, en consecuencia, son poco flexibles para recoger los cambios de estructurales, y (ii) no tienen definidas las operaciones de conjuntos necesarias para manipular las categorías.

El componente sintáctico del modelo higraph representa, de forma natural, la compleja naturaleza estructural de los tesauros, formada por categorías que agrupan términos y

relaciones, binarias y n-arias, entre los términos y/o categorías; además, los cambios permanentes que afectan a la estructura o al contenido de los tesauros pueden representarse con las operaciones de conjuntos y las operaciones de grafos.

Los modelos de implementación de datos que hemos analizado son, por un lado, el modelo relacional, el modelo más extendido en la construcción de sistemas informáticos de bases de datos y, por otro lado, los modelos basados en los lenguajes de marcado XML, más utilizados en la representación de información en la Web. Aunque ambos tipos de modelos son de naturaleza muy diferente, los primeros se basan en estructuras de datos planas y los segundos en estructuras jerárquicas; existen, además, algoritmos para transformar, automáticamente, unos en otros. La ventaja del modelo relacional es que se gestiona la información de forma más eficaz, pero los modelos basados en XML son los que mejor resuelven el intercambio, compartición y reutilización de la información.

El modelo relacional organiza la información en estructuras matemáticas denominadas relaciones. Las relaciones, en matemáticas, tienen un significado diferente de las relaciones entendidas como asociaciones. La relación matemática es un conjunto de tuplas, y una tupla es una lista ordenada de valores; las relaciones se suelen visualizar con tablas, donde cada fila es una tupla, y cada columna es un atributo de la tupla. Las relaciones matemáticas sirven para representar tanto las entidades como las relaciones de tipo asociación; en el primer caso, cada tupla es una entidad y cada valor de la tupla es el valor de un atributo de esa entidad; en el segundo caso, cada asociación es una tupla que contiene todas las entidades participantes en esa asociación. El modelo relacional, además, proporciona un conjunto de operaciones y un lenguaje declarativo para expresar las operaciones. Los Sistemas de Gestión de Bases de Datos relacionales almacenan y gestionan automáticamente grandes cantidades de datos de forma eficiente y segura si el diseño de la base de datos es adecuado. Los diseños son adecuados si cumplen dos requisitos: (i) evitar al máximo la repetición de los mismos datos, y (ii) ser exhaustivos. Los diseños relacionales de tesauros realizados hasta ahora se caracterizan porque reproducen el esquema sistemático del tesauro; el inconveniente de estas aproximaciones es, como en los modelos conceptuales, que el esquema de datos es fijo y no se puede cambiar una vez construida la base de datos relacional. Por lo tanto, si depende del esquema sistemático del tesauro obtenido en un determinado momento, cuando el ámbito del tesauro cambia, el tesauro relacional se queda obsoleto y hay que reconstruirlo. Otro inconveniente es que el contenido de los tesauros es dependiente del

esquema de datos relacional por lo que es difícil intercambiar, compartir y reutilizar este contenido en otros tesauros con esquemas de datos diferentes.

Con el objeto de mejorar la interoperabilidad entre tesauros y sistemas y de representar y manipular de forma más natural las relaciones jerárquicas, se utiliza un segundo tipo de modelo de implementación de datos basado en el metalenguaje estándar XML. Los modelos XML son lenguajes de marcado que se crean utilizando el metalenguaje estándar de la Web, XML. Estos lenguajes se definen en archivos de texto mediante gramáticas independientes de contexto que pueden estar aumentadas con pares atributo-valor denominados DTD o Esquemas. Los tesauros que se construyen con estos lenguajes son archivos de texto en los que el contenido está estructurado e interpretado con etiquetas o marcas XML. En nuestro análisis hemos distinguido tres tipos de lenguajes: 1) los que están basados en RDF, 2) los que provienen del *e-learning*, y 3) los que se han definido libremente.

Los primeros comparten un metamodelo que es estándar en la Web, el RDF, lo que, en teoría, proporciona mayor interoperabilidad. Este metamodelo está basado en describir los contenidos con ternas (objeto, propiedad, valor). RDF se ha utilizado para definir, entre otros muchos lenguajes, los estándares RDFS, OWL y SKOS-Core. El RDFS es una extensión del RDF para definir la semántica de los elementos de RDF y para crear clases, que son tipos de objetos que comparten propiedades. El OWL es un modelo que amplía la semántica básica de RDFS con una semántica formal, proporcionando una capacidad de abstracción mayor con el objetivo de representar y procesar conocimiento; es el modelo o lenguaje de representación de las ontologías en la Web. El SKOS-Core es una simplificación del complejo modelo OWL que está orientado específicamente a la representación de tesauros, pero concebidos como ontologías terminológicas, es decir, como un conjunto de conceptos relacionados semánticamente entre sí y relacionados con términos del lenguaje natural.

El segundo tipo de lenguajes XML que hemos distinguido son los lenguajes IMS VDEX y CEN XVD definidos para los vocabularios de explotación del *e-learning*. El IMS Vocabulary Definition Exchange, IMS VDEX, es el modelo de representación e intercambio de vocabularios, listas, taxonomías y tesauros, utilizados en aplicaciones *e-learning*; es un modelo con una capacidad expresiva limitada porque representa el contenido de los tesauros únicamente con términos y relaciones, no considera categorías y utiliza un esquema de datos fijo. El modelo CEN Exchange of Vocabularios, CEN XVD, consta en realidad de un modelo para construir tesauros multilingües y un modelo

para definir las correspondencias entre vocabularios. El modelo de contenido consta de dos tipos de estructura: la primera sirve para agrupar todos los términos equivalentes, y la segunda se utiliza para definir el resto de relaciones semánticas entre términos. En cualquiera de los dos casos, estos modelos reproducen los modelos de contenido de los objetos de aprendizaje *e-learning*, lo que significa que el tesauro se concibe como un objeto de aprendizaje más, que puede ser manipulado de la misma forma por las aplicaciones *e-learning*.

Finalmente los lenguajes no estándar, definidos libremente de acuerdo con los esquemas sistemáticos de los tesauros, son lenguajes creados ad hoc, que tienen la ventaja de representar fielmente la estructura y propiedades de cada tesauro particular, pero que aportan poco a la resolución del problema de la interoperabilidad.

Los modelos XML surgieron, en realidad, para mejorar la interoperabilidad, pero en su aplicación real se han creado innumerables lenguajes, estándares y particulares, resultando que, de nuevo, es muy complicado el poder compartir y reutilizar los tesauros, al igual que ocurre con los esquemas de datos diferentes generados con el modelo relacional. Por eso, actualmente, la solución que se adopta mayoritariamente es utilizar el modelo relacional, más eficiente y seguro, para construir los tesauros, y el modelo XML para definir el contenido y el esquema relacional del contenido en las operaciones de intercambio de información entre los sistemas que operan con los tesauros.

9.1.4.5. Métodos de construcción de tesauros

Finalmente, en el análisis de los métodos de construcción de tesauros monolingües hemos distinguido entre métodos deductivos e inductivos. Ambas aproximaciones están basadas en los estándares de construcción de tesauros y pueden aplicarse a la construcción manual, semiautomática o automática de los tesauros. El proceso de construcción de tesauros es un proceso empírico e iterativo. Es empírico porque está sujeto al ámbito concreto del tesauro: su esquema conceptual y su terminología. Es iterativo porque repite un conjunto de operaciones hasta que se completa el proceso. Distinguimos siete operaciones básicas en la construcción de los tesauros, que en los procesos de tipo inductivo siguen el orden siguiente:

(i) el análisis del dominio: son los preliminares donde se establece el objetivo y naturaleza del tesauro en un documento que se denomina plan o planta del tesauro;

(ii) selección de fuentes: consiste en definir los textos que, de acuerdo al plan, servirán para extraer los términos del futuro tesauro; pueden ser fuentes estructuradas o no;

(iii) recolección de términos: consiste en seleccionar, de las fuentes, los términos con mayor contenido semántico;

(iv) análisis y clasificación: consiste en preparar un esquema conceptual sistemático para el tesauro; en los métodos inductivos se prepara con los términos que se ha recogido, aplicando técnicas de agrupación semántica y luego estableciendo las relaciones entre los términos; en las metodologías deductivas primero se establece el esquema conceptual del dominio, con ayuda de expertos de cada una de las disciplinas que abarca y aplicando técnicas de análisis por temas o por facetas; el resultado en ambos casos es el esqueleto del tesauro, la macroestructura, que servirá de base para la implementación;

(v) implementación: es la construcción del tesauro en algún soporte, papel o electrónico; los tesauros de explotación se construyen como los sistemas informáticos: primero, se diseña el esquema de datos aplicando un modelo de implementación de datos que, normalmente, reproduce el esquema conceptual obtenido en la operación anterior; y, después, se insertan las instancias recolectadas de las fuentes, términos, relaciones y categorías, hasta completar el tesauro;

(vi) publicación: consiste en hacer accesible el contenido del tesauro; y

(vii) uso, evaluación y mantenimiento: estas tres operaciones deben realizarse de forma permanente durante la vida útil del tesauro; el uso servirá para detectar errores y omisiones por lo que contribuye a la evaluación y al mantenimiento; la evaluación comprueba cuatro aspectos: a) la corrección formal del tesauro, si incluye los preliminares, si son correctas las formas de presentación, si se ajusta al dominio, si los términos son suficientemente descriptivos, etc.; b) la riqueza relacional del tesauro, son medidas cuantitativas sobre el grado de conexión, la tasa de enriquecimiento, la conectividad, etc.; c) la usabilidad, que mide lo útil que es el tesauro para el usuario; d) la eficiencia para la RI, comparando la ganancia en exhaustividad y precisión obtenida con el tesauro. El mantenimiento, finamente, es la etapa más delicada en un tesauro porque el dominio y la lengua están en

> permanente cambio y el tesauro, para ser efectivo, debería recoger estos cambios en el momento en que surgen; también porque pueden afectar a la consistencia de la información o del esquema de datos sobre el que se construye el tesauro. Para llevar a cabo el mantenimiento de forma controlada es necesario definir una metodología, que suele consistir en ir recogiendo las modificaciones, términos obsoletos, nuevos términos, cambio de estatus de los términos, etc., de forma sistemática y, en períodos de tiempo determinados, estas modificaciones, una vez analizadas, se ejecutan o se rechazan. El problema es que las modificaciones son tan costosas que o no se hacen por falta de medios o se hacen en periodos de tiempo largos, normalmente años, con lo que el tesauro nunca está totalmente actualizado. Esto, en determinados dominios muy cambiantes, hace que el tesauro sea, en poco tiempo, inservible; por ejemplo, en los dominios de información tecnológicos o en colecciones de recursos digitales que están en cambio permanente.

En las metodologías deductivas, el análisis y diseño conceptual se realiza en segundo lugar, de forma que el orden de las operaciones es el siguiente: (i) análisis del dominio, (ii) análisis y diseño conceptual, (iii) selección de fuentes, (iv) recolección clasificada de términos, (v) implementación, (vi) publicación, y (vii) uso, evaluación y mantenimiento. En las metodologías inductivas, las etapas de recolección de términos, análisis y clasificación se repiten para ir refinando tanto la lista de términos como el esquema conceptual del tesauro; igualmente, en las metodologías deductivas las operaciones de análisis y diseño conceptual, selección de fuentes y recolección de términos se repiten hasta tener completo el esquema y la lista de términos. También la etapa de uso, evaluación y mantenimiento conlleva la repetición de las operaciones anteriores porque, en el mejor de los casos, durante el uso y la evaluación se van a seguir recolectando e insertando términos en el tesauro para mantenerlo actualizado, y, en el caso peor, los nuevos términos pueden obligar a volver al análisis y diseño conceptual, para modificar el esquema del tesauro y repetir el resto de operaciones hasta la publicación. Finalmente, algunas de las operaciones pueden automatizarse: (i) la selección de fuentes, comparando el contenido de éstas con el ámbito del tesauro; (ii) la recolección de los términos, aplicando técnicas de procesamiento de lenguaje natural o técnicas estadísticas para analizar los textos y extraer los términos con más significado; y (iii) el análisis y clasificación, para obtener el esquema conceptual del tesauro

mediante técnicas estadísticas que agrupan los términos con significados cercanos. En cualquier caso, la construcción automática o semiautomática de tesauros tiene que completarse con la intervención humana si se quieren obtener tesauros de la misma calidad que los tesauros construidos manualmente.

9.1.5. Conclusiones parciales del análisis

De cada uno de los análisis hemos extraído una serie de conclusiones parciales que nos han servido para diseñar la metodología que completa la demostración de la hipótesis. Respecto del primer análisis, sobre la naturaleza y aplicaciones de los tesauros de explotación en el contexto *e-learnig*, hemos concluido que los vocabularios de explotación, para poder se eficazmente utilizados por los equipos docentes, deberían: a) ajustarse al ámbito de uso, que son los contenidos educativos y las colecciones de recursos didácticos o de investigación; b) ajustarse al lenguaje propio de la comunidad académica que utiliza el vocabulario y que crea el dominio de aplicación del tesauro; c) servir de marco conceptual de referencia del conocimiento y trabajo de cada comunidad docente, discente de investigación específica; y d) servir para enseñar y aprender sobre el dominio de conocimiento y el lenguaje de especialidad que es el ámbito del tesauro.

Respecto del segundo análisis, sobre el contexto actual de trabajo académico *e-learning* donde emergen los tesauros académicos de explotación, hemos concluido que en este momento la tecnología *e-learning* es el catalizador tecnológico para la enseñanza y aprendizaje universitarios pero, para asegurar que realmente apoyen una enseñanza, aprendizaje e investigación de calidad, es necesario resolver algunas cuestiones de las que destacamos: a) desde el punto de vista institucional, el reconocimiento y valoración del trabajo del profesor en los entornos virtuales; b) el reconocimiento de las creaciones didácticas e investigadoras publicadas en los entornos virtuales como parte de la producción científica; c) la definición de modelos y métodos que faciliten la creación de espacios de aprendizaje de calidad; y d) la definición de modelos y métodos que faciliten al profesor la construcción y explotación de materiales y recursos didácticos de calidad.

Respecto del tercer análisis, sobre las estructuras-t que sirven de fuente a los tesauros académicos de explotación, hemos concluido que son pequeñas redes de términos relacionados semánticamente, construidos por los profesores en los entornos virtuales, con su conocimiento del dominio y del lenguaje de especialidad, para clasificar y describir los sus creaciones intelectuales, los materiales y recursos digitalizados

didácticos y de investigación. Constituyen una forma simplificada de definir y clasificar el contenido de estos recursos en las aplicaciones informáticas que los gestionan como, por ejemplo, los repositorios de objetos de aprendizaje. En estos repositorios, las estructuras-t se localizan en los metadatos de los recursos y se pueden reutilizar para construir, con ellas, un tesauro del repositorio.

Respecto al cuarto análisis, sobre la evaluación de los modelos utilizados para construir tesauros, hemos concluido lo siguiente: a) el modelo de los estándares de construcción de tesauros de explotación aporta un marco común para denotar los elementos del contenido de los tesauros, los modos de presentación y las reglas de modificación, por lo cual, es un referente, imprescindible, en el diseño de los tesauros; b) de los modelos de datos analizados a nivel conceptual, el modelo higraph es el único que representa de forma natural las estructuras léxicas de los tesauros, aunque es necesario definir una semántica que sea compatible con el modelo semántico de los higraphs y con la concepción del tesauro como sistema estructurado de signos en el que el valor del significado de cada signo depende de su posición; y c) de los modelos de implementación de datos analizados, el modelo relacional es el más adecuado para implementar los tesauros de explotación porque es el más eficiente desde el punto de vista informático y, respecto a la interoperabilidad, teniendo en cuenta que no existe un único modelo XML estándar, ésta está parcialmente resuelta con la opción que tienen los sistemas actuales de gestión de bases de datos relacionales de exportar automáticamente el contenido a documentos XML y aplicar, después, correspondencias con los modelos XML de las aplicaciones con las que se desea interoperar.

Finalmente, respecto del quinto análisis sobre los métodos de construcción de los tesauros de explotación, hemos concluido que: a) es un proceso largo y costoso que puede aliviarse si se automatizan algunas de las siete operaciones, pero para obtener tesauros de calidad, es necesaria la intervención de expertos; b) la construcción, en realidad, es un proceso continuado para que el tesauro no se quede obsoleto y pierda eficacia; y c) la etapa más delicada es el mantenimiento, porque si no se realiza metodológicamente puede afectar negativamente a la coherencia del contenido y, en ciertos casos, cuando afecta al esquema de datos, inevitablemente supone rehacer el tesauro.

Estas conclusiones nos han permitido verificar los supuestos de la hipótesis (i) a (iv): (i) el tesauro es un sistema estructurado de signos lingüísticos en el que el valor del significado de cada signo depende de su posición diferencial respecto de los demás; (ii)

el tesauro representa las nociones de un dominio de conocimiento mediante signos organizados en grupos por los tipos de relaciones semánticas de generalización-especialización, equivalencia, asociación y, posiblemente, otras relaciones específicas del dominio; (iii) existe un modelo formal único para representar esta concepción de tesauro; y (iv) existen estructuras de términos en semántica libre que representan parcialmente, las nociones de un dominio mediante signos organizados en grupos por las relaciones semánticas mencionadas y que son creadas libremente por los profesores para describir y clasificar sus creaciones intelectuales. Asimismo, nos permite responder parcialmente a las cuestiones de investigación que planteamos.

Respecto a la primera cuestión, cómo definir los tesauros académicos de explotación de forma que su contenido se ajuste al dominio de conocimiento, al lenguaje y a las necesidades concretas del los profesores, consideramos que una posible respuesta podía ser que estos tesauros fueran el resultado de recoger organizadamente las estructuras-t, creadas por los profesores para clasificar los materiales didácticos o de investigación del dominio. Para terminar de responder a la cuestión, sin embargo, nos quedaba encontrar o crear ejemplos que ilustrasen esta definición y encontrar o crear el mecanismo de organización de estas estructuras-t, lo cual era parte de la segunda cuestión de investigación.

Respecto de la segunda cuestión de investigación, cómo definir los mecanismos que permitan a los usuarios construir y mantener actualizados estos tesauros de la forma más sencilla posible, considerábamos que el modelo matemático higraph podía ser el metamodelo más adecuado para representar cualquier tipo de contenido de los tesauros, siempre y cuando fuera posible extender el modelo semántico para dar, a los elementos estructurales del modelo higraph, el significado adecuado a nuestro concepto y uso de tesauro y, además, fuera posible encontrar un modelo de implementación de datos que represente, con exactitud, los higraphs conceptuales. Además, considerábamos que el método de construcción y mantenimiento debía ser, necesariamente, inductivo, puesto que, de acuerdo con las conclusiones, tenía que recoger las estructuras-t para construir este nuevo tipo de tesauro académico de explotación.

9.1.6. Método de demostración

Sabiendo qué es lo que nos faltaba para completar las respuestas a las cuestiones de investigación y cuáles eran las conclusiones, formuladas en la hipótesis, que podrían verificarse, diseñamos y aplicamos el siguiente método de demostración: 1) selección y

recogida de datos: las fuentes de estructuras-t y los ejemplos de tesauros o vocabularios académicos de explotación; 2) aplicación del modelo sintáctico del higraph a los datos; 3) desarrollo de la extensión semántica de los higraphs; 4) experimentación del modelo higraph, completo, con los datos; 5) sistematización del método de construcción aplicado manualmente en 2 y 4; 6) experimentación con los datos del método de construcción; y 7) evaluación de resultados y conclusiones.

9.1.7. Recogida de datos

Los datos que nos interesaban eran ejemplos de a) estructuras-t, y b) tesauros académicos de explotación o, en su defecto, cualquier otro tipo de vocabulario que sirviera para demostrar la definición de tesauro académico de explotación y para probar la viabilidad del modelo extendido de higraph y el método de construcción. Estos datos los recogimos de los proyectos de investigación en los que hemos participado[2], aunque también encontramos ejemplos interesantes, elaborados por nuestros estudiantes[3], o por otros colegas[4], en las asignaturas del campus virtual UCM. Seleccionamos tres ejemplos concretos: 1) la adaptación del tesauro europeo ETB a las nociones impartidas en la enseñanza primaria y secundaria de España, que fue realizada por los grupos de trabajo GT9, GT8 de AENOR, para su utilización en el repositorio educativo AGREGA, y en el que participó uno de los componentes del grupo de investigación e-UCM de la Facultad de Informática UCM; 2) la construcción de un índice de términos de arqueología precolombina del museo virtual CHASQUI, realizado a partir de las estructuras-t de los metadatos LOM que clasifican los objetos virtuales del museo y que han sido creadas por el grupo de investigación de la Facultad de Geografía e Historia, participante en el proyecto OdA; y 3) la construcción del tesauro e-Derecho, realizado a partir de las voces relacionadas de un glosario explicativo creado por expertos del derecho y de las tecnologías de la información y las comunicaciones y por el equipo de investigación de lingüística LALINGAP, en el marco del proyecto de innovación PIMCD 66/2008.

En un primer experimento, utilizamos el modelo higraph para estructurar y dibujar el contenido de los vocabularios que habíamos escogido como ejemplos. El resultado fue los higraphs que mostraban las estructuras de estos tres vocabularios, incluso, como en

[2] Ver capítulo 1.
[3] Por ejemplo, las clasificaciones de los corpus creados por los estudiantes de Análisis del Discurso.
[4] Por ejemplo, los vocabularios para clasificar bancos de imágenes.

el caso del higraph de CHASQUI, se podían observar estructuras que no estaban explícitas en el índice original. Para construir estos higraphs experimentales utilizamos una correspondencia entre los elementos del higraph y los elementos estándar de los tesauros, por ejemplo, los términos se representaban con los nodos atómicos, las categorías con los nodos no atómicos y las relaciones semánticas TG/TE y USE/USEP con arcos y las relaciones de asociación TR con hiperarcos. Esta correspondencia sirvió de base para desarrollar la extensión del modelo higraph al dominio léxico, modelo HL, que proponemos para el diseño conceptual de los tesauros. En un segundo experimento, utilizamos el modelo higraph para ir integrando, incrementalmente, un conjunto seleccionado de estructuras-t con las que se habían creado los vocabularios ejemplo. El resultado fue que, de nuevo, obteníamos los mismos higraphs que habíamos construido en el primer experimento y que el proceso de construcción podía realizarse siguiendo una serie de pasos que nos dieron la pauta para diseñar el método inductivo HL que proponemos.

9.1.8. El modelo propuesto

El modelo que proponemos es una aplicación del modelo higraphs al dominio léxico, por eso lo hemos denominado modelo higraph léxico o, abreviadamente, HL. Este modelo tiene, como los higraphs, dos componentes: sintáctico y semántico. El componente sintáctico está formada por cuatro elementos: (i) un conjunto de signos lingüísticos, que son los términos y categorías del tesauro, (ii) una función de inclusión, σ, que define el contenido de las categorías del tesauro y establece que los términos son los nodos atómicos del higraph, es decir, que son nodos que no contienen nada; (iii) una función de partición, π, que define cómo se divide una categoría en clases de equivalencia disjuntas, por ejemplo, en facetas; y (iv) un conjunto de relaciones, que son las relaciones semánticas entre los signos del tesauro. Además, en el HL están definidas las operaciones de conjuntos, unión, intersección, diferencia, producto cartesiano y las operaciones de grafos, búsqueda de nodos, cálculo de caminos, extracción de jerarquías y redes de un mismo tipo, inserción, borrado y modificación de nodos o de relaciones, etc., que definen la componente funcional del sistema HL.

El componente semántico del HL es una extensión del modelo semántico de los higraphs que tiene el propósito de incluir el concepto del valor del significado de los signos de un tesauro, entendido éste como un sistema lingüístico sujeto al principio de solidaridad. Así, la interpretación μ de un término es el valor de su significado y

depende de su posición diferencial respecto de los demás signos del HL. Formalmente se calcula como el conjunto de relaciones semánticas que mantiene con los otros signos del HL, y cada relación semántica se expresa con una terna (término1, relación semántica, término2). La interpretación de una categoría es, también, su valor semántico y se calcula, aplicando el modelo semántico de los higraph, con la composición incremental de los significados de los términos y categorías que contiene. Formalmente esta composición se realiza con las operaciones de unión y producto cartesiano del conjunto de los valores semánticos de los componentes. En definitiva, el modelo semántico HL formaliza el proceso de obtener el significado de un signo que, de forma intuitiva, realizamos las personas cuando leemos un tesauro: creamos el significado de un signo con los significados de los signos más cercanos, valorando si el signo cercano es más general, es equivalente, o está asociado.

Para poder experimentar la viabilidad y las posibilidades del modelo HL es necesario implementarlo en algún sistema informático, utilizando herramientas que permitan la creación y manipulación automática del contenido. Utilizamos el modelo relacional porque: 1) no existen, evidentemente, herramientas software para gestionar HL y aunque sí existen herramientas para gestionar higraphs, éstas no son adecuadas para los HL porque a) o bien no se pueden aplicar porque son de propósito específico, o bien b) las que conocemos de propósito general son muy limitadas funcionalmente, porque sólo permiten dibujar y explorar el higraph, son, además, costosas de utilizar, porque utilizan un lenguaje propio de definición que dificulta la interoperabilidad con otras herramientas de gestión; 2) del análisis de los modelos de implementación de datos concluimos que el modelo relacional es eficiente, dispone de múltiples aplicaciones de gestión y tiene parcialmente resuelto el problema de la interoperabilidad con otros sistemas; y 3) las estructuras de representación de datos del modelo relacional, las relaciones matemáticas, coinciden con las estructuras de representación semántica del HL, lo que facilita enormemente la implementación, puesto que basta con almacenar directamente los valores del significado de los signos del HL.

El esquema relacional del HL que hemos desarrollado está formado por dos tipos de relaciones matemáticas básicas que hemos llamado hl_macro y hl_micro. El primero, hl_macro, almacena los valores del significado de todos los signos del HL, es decir, la interpretación de la macroestructura, que están formadas por las ternas (tipo de relación, signo1, signo2), pero con una serie de restricciones que tienen el objetivo de (i) reducir al mínimo la redundancia de los datos, (ii) mantener la coherencia y, finalmente, (iii) no

perder información. El segundo tipo de relación, hl_micro almacena todos los signos del HL con la indicación del tipo de signo, término o categoría; esta relación es, en realidad, redundante, puesto que los signos podrían extraerse de hl_macro, pero facilita la gestión del HL y ayuda a mantener la consistencia de los datos.

El modelo HL es la respuesta que damos a la segunda cuestión de investigación, encontrar un metamodelo o modelo general para representar de forma sistemática el contenido de tesauros de cualquier tipo y, también, demuestra los enunciados (i) y (iii) de la hipótesis de investigación: (i) considerando que el tesauro es un sistema estructurado de signos lingüísticos en el que el valor del significado de cada signo depende de su posición diferencial respecto de los demás; (iii) existe un modelo formal único para representar esta concepción de tesauro.

9.1.9. El método propuesto

El método que proponemos es una adaptación del método general inductivo de construcción de tesauros basada en 1) utilizar como fuente las estructuras-t, y 2) utilizar el modelo HL para integrarlas en un único sistema. Simplifica el proceso de construcción porque se prescinde de las primeras etapas del proceso, el análisis del dominio, la selección de fuentes, la recolección de términos y el análisis y la clasificación de los mismos. El análisis del dominio, en el caso concreto de los tesauros académicos de explotación, consiste, simplemente, en establecer la colección de recursos o los materiales que van a constituir el ámbito del tesauro. El resto de los parámetros, como el propósito o el tipo de usuario, ya están definidos por su carácter académico y de explotación. La selección de fuentes no es necesaria, porque son las estructuras-t contenidas en los materiales o colecciones de recursos pero, en cambio, es necesaria la intervención del equipo docente, autor de las estructuras-t, para su definición e identificación. La recolección de términos, el análisis y la clasificación se resumen en las fases, automáticas, de extracción y análisis HL de las estructuras-t. El resto de las fases, inserción en la base de datos, publicación, uso y validación son comunes. La fase de actualización y mantenimiento, sin embargo, es una continuación de la construcción, y se realiza dinámicamente cada vez que hay nuevas estructuras-t. El método HL simplifica el proceso de construcción de tesauros académicos de explotación, pero sólo puede aplicarse si se cumplen las premisas de: (i) disponer de contenidos o recursos didácticos y/o de investigación, creados y documentados por los

equipos de profesores que necesitan el tesauro o, al menos, que se hayan creado utilizando el mismo lenguaje, y (ii) que contengan estructuras-t.

El proceso comienza a partir de una colección de recursos o de materiales para los que se construye el tesauro, y en los que se han identificado un mismo tipo de estructuras-t; consta de seis fases: 1) el usuario identifica y define el tipo de estructuras-t que contiene su colección; 2) se extraen, automáticamente, las estructuras-t utilizando la definición de tipo de estructura-t de la fase anterior; 3) se analizan e interpretan las estructuras-t para obtener los componentes que se van a insertar en el HL del tesauro, utilizando la definición de tipo de estructura-t y el modelo HL; 4) se revisan y adecuan estos componentes, de forma manual, para aplicar las reglas de control y/o depurar el tesauro, y 5) se insertan, automáticamente, los componentes HL en la base de datos relacional HL. Las etapas 2 a 5 se repiten hasta que no queden nuevas estructuras-t. La etapa de inserción se apoya en el modelo HL para incorporar, de forma incremental, las estructuras-t en un esquema HL que inicialmente está vacío. En la fase de análisis e interpretación de las estructuras-t se aplica una estrategia basada en la construcción automática de series semánticas, que son series de signos relacionados por un mismo tipo de relación, a partir de la definición del tipo de estructura-t que proporciona el usuario. Además, se aplica una estrategia basada en la frecuencia de aparición de las estructuras-t para evitar que las posibles estructuras-t erróneas o contradictorias generen inconsistencias. El proceso de construcción puede automatizarse, excepto (i) la fase primera, identificación y definición de estructuras-t, y (ii) la cuarta, revisión y adecuación, que requiere de la intervención del usuario. Es un proceso inductivo que puede ejecutarse siempre que existan estructuras-t, con lo que la construcción y mantenimiento de tesauros se funden en una misma operación. El esquema conceptual del tesauro es un HL que va surgiendo y va cambiando conforme se insertan las estructuras-t en el esquema de datos relacional HL que, en realidad, puede considerarse un metaesquema para crear los HL.

Con el método HL se responde a la última de las cuestiones de investigación y, respecto de la hipótesis, demuestra que es posible construir y actualizar, de forma sistemática, los tesauros académicos de explotación a partir de las estructuras-t creadas por los equipos docentes para describir y clasificar los materiales.

9.1.10. La experimentación

La última parte de esta investigación la hemos dedicado a experimentar, con los casos seleccionados, la viabilidad, ventajas, e inconvenientes del modelo y método HL. Estos casos prácticos proceden de desarrollos reales que se hicieron sin aplicar las propuestas de este trabajo de tesis, pero que han servido para motivar, apoyar, y verificar esta investigación. Hemos realizado un proceso de ingeniería inversa para reconstruir cada uno de los vocabularios como un tesauro aplicando el modelo y método HL. Los casos elegidos muestran la utilidad del método para resolver diferentes tipos de necesidades que han surgido o que pueden surgir en el contexto del *e-learning*.

9.1.10.1. La especialización de tesauros generales

El primer caso práctico consiste en aplicar el método HL para especializar tesauros generales al dominio de las colecciones de recursos educativos. Esta especialización se realiza, normalmente, de forma deductiva y manual. Uno de los casos más recientes es la especialización del tesauro Europeo ETB a los contenidos de los recursos educativos de educación primaria y secundaria en España realizada por un equipo de expertos de AENOR. Nuestra propuesta pretende aportar un método alternativo que facilite este proceso a los equipos de profesores en un ámbito de aplicación restringido a sus contenidos didácticos o de investigación. Para ello teníamos que (i) demostrar que la propuesta era viable, y (ii) encontrar cuáles eran las ventajas e inconvenientes frente a una aproximación deductiva tradicional. Para llevar a cabo este caso práctico utilizamos de muestra la subcategoría 70.90, 'Ética/Religión/Ideología', del tesauro ETB en español, y el lenguaje LOM-ES de los metadatos de los objetos de aprendizaje del repositorio AGREGA. Además, construimos una muestra de estructuras-t con los términos y relaciones nuevas que el equipo de AENOR había introducido o modificado en el tesauro ETB; de esta forma simulamos que los cambios procedían de nuevas clasificaciones de supuestos recursos del repositorio de objetos de aprendizaje, es decir, el caso en el que se encontraría un equipo docente que necesitase especializar un tesauro general a su colección de recursos educativos. Una vez construidas las estructuras-t, aplicamos el método HL para insertarlas en la base de datos relacional HL que contenía la información inicial del tesauro ETB sin las correcciones de AENOR. El resultado fue (i) que el método HL era viable, y comprobamos que (ii) la base de datos relacional HL recogía sin problemas las nuevas estructuras; (iii) que la estructura del HL del tesauro cambiaba dinámicamente según se iban incorporando nuevas estructuras-t; y (iv) que en

todo momento podíamos consultar el estado del tesauro, utilizando el lenguaje de consulta del gestor. Sin embargo, también detectamos algunos problemas que no habíamos tenido en cuenta entre las que destacan: (v) que era imprescindible realizar la etapa de revisión y adecuación para evitar la proliferación de estructuras-t erróneas e incompatibles con las del tesauro, como consecuencia de la estrategia de resolución de inconsistencias del método HL, que no borraba las estructuras-t inconsistentes. Dedujimos que es un mecanismo de selección que funciona bien si el índice de errores es bajo pero que, cuando el conjunto de estructuras-t inservibles crece a lo largo del tiempo, el tesauro es menos eficiente; (vi) que la calidad del tesauro resultado depende de la calidad de las nuevas estructuras-t, de la fase de revisión y corrección manual para eliminar, o aceptar, las estructuras, términos, categorías y relaciones, que eran incompatibles. En todo caso, comprobamos que el método HL, frente al método deductivo tradicional, es más fácil de aplicar porque (a) reutiliza estructuras-t ya construidas que son la fuente de especialización; y (b) no es necesario revisar el esquema conceptual ni el contenido del tesauro porque éste se va adaptando inductivamente.

9.1.10.2. La reconstrucción, como tesauro, del índice temático de un museo virtual académico

El segundo caso práctico se refiere a la construcción de un tesauro académico de explotación del repositorio académico de objetos virtuales CHASQUI. En este caso se trataba de reconstruir el índice temático del repositorio de objetos virtuales del museo académico CHASQUI como un tesauro para (i) comprobar, de nuevo, las viabilidad y posibilidades del método HL en un nuevo problema: construir inductivamente tesauros de repositorios de recursos didácticos o de investigación digitalizados, y (ii) entender la estructura y significado del vocabulario que contenía el índice temático del repositorio construido a partir de las estructuras-t sin aplicar el modelo y método HL. Estos datos sirven para mejorar la eficacia del índice o para construir un tesauro de explotación que muestre la estructura terminológica y conceptual del dominio de conocimiento del repositorio. El procedimiento empleado fue aplicar el método HL a las estructuras-t extraídas directamente de los metadatos de los objetos virtuales del repositorio CHASQUI. Estos metadatos, al igual que ocurre en el repositorio AGREGA del caso 1, están basados en el estándar LOM, pero el equipo docente de CHASQUI los interpreta y utiliza de forma diferente para poder añadir más tipos de relaciones. Este nuevo uso de

LOM puso en evidencia la necesidad de llevar a cabo de forma manual la primera etapa del método HL, la identificación de estructuras-t, solicitando al usuario que definiera cómo había interpretado y utilizado las estructuras-t. Esta definición se realiza siempre mediante una tabla de correspondencias en la que el usuario describe los componentes de una estructura-t tipo, utilizando un metalenguaje simple y cercano al estándar de tesauros y al estándar XML, y las marcas textuales que sirven para identificarlos en las fuentes de estructuras-t. Los resultados nos han servido para comprobar que el modelo HL permite (i) sistematizar el contenido del vocabulario implícito en todas las clasificaciones de los objetos virtuales del repositorio; (ii) hacer explícitas las estructuras semánticas contenidas implícitamente en el índice, incluso aquellas que no eran evidentes; (iii) se puede representar gráficamente la estructura del tesauro de CHASQUI; (iv) se puede construir el tesauro de CHASQUI, inductivamente, a partir de las clasificaciones de los objetos del repositorio; (v) se puede actualizar dinámicamente el tesauro aplicando el método HL cuando existan nuevas estructuras-t; (vi) el tesauro CHASQUI representa fielmente las nociones del dominio de conocimiento de los objetos virtuales, puesto que recoge las estructuras-t que han sido cuidadosamente creadas o seleccionadas por los profesores para describir los objetos del repositorio; (vii) el método HL simplifica la creación del tesauro, porque se sustituyen las costosas operaciones de planificación, selección de fuentes y análisis y clasificación por la definición, manual, del tipo de estructura-t, la extracción y análisis automático de dichas estructuras-t; (viii) la etapa de revisión y adaptación se puede llevar a cabo durante el uso del tesauro, tal y como se hizo, de forma satisfactoria, en el índice; además, esta etapa no sólo sirve para depurar el tesauro sino, también, los objetos virtuales del repositorio asociados a los términos; y, finalmente, (ix) la integración del tesauro HL en el repositorio es más sencilla si el esquema de implementación de datos del HL debe utilizar el mismo modelo de de datos que el repositorio.

9.1.10.3. La creación de un tesauro para el Glosario *explicativo e-Derecho*

El último caso práctico trata sobre la creación de un tesauro académico de explotación del Glosario *e-Derecho* sobre el ámbito del Derecho y la propiedad intelectual en Internet. Presenta dos diferencias significativas con respecto de los casos tratados anteriormente: (i) las estructuras-t no han sido extraídas de los metadatos, sino del contenido del glosario, y (ii) el objetivo de este caso es, además de explotar el contenido del glosario, sistematizarlo. El Glosario *e-Derecho* fue creado de forma libre y

cooperativa por un equipo mixto de docentes y profesionales expertos del Derecho y de las Tecnologías de la Información y de las Comunicaciones. Contiene un conocimiento rico, interdisciplinar que es difícil de encontrar en otras obras, pero carece de un modelo de contenido que sistematice conceptualmente todo ese conocimiento, de forma que sea más aprovechable para las personas o aplicaciones informáticas que lo explotan. En consecuencia, este experimento nos permitía comprobar la flexibilidad del modelo y método HL al aplicarlo a un tipo de estructuras-t nuevas y a un tipo de explotación que no habíamos previsto: la sistematización de creaciones intelectuales creadas libremente, sin responder a estructuras previas.

El procedimiento para aplicar el método HL fue el siguiente: a) se identificaron y definieron de la forma prevista en el método, mediante la tabla de correspondencias, las estructuras-t que estaban etiquetadas en el contenido del glosario explicativo; (ii) se creó la base de datos relacional HL; (iii) se aplicó el método, sin modificar los procesos de extracción, análisis e interpretación utilizados en los dos casos anteriores. El resultado obtenido fue el esperado y resultó posible: (i) comprobar la flexibilidad del modelo y método HL, y (ii) obtener un tesauro, con todos los términos del glosario, pero organizados en un HL que sistematiza y reproduce las estructuras de términos relacionados que los autores del glosario habían incluido en el material original. Se puede considerar que el tesauro es una interfaz de acceso, superpuesta al glosario, que ayuda al usuario a localizar las nociones que se explican en él.

9.1.10.4. Evaluación de los tesauros resultado

No es sencillo evaluar la usabilidad y efectividad de los tesauros construidos dentro de este trabajo de investigación, un prototipo y un tesauro real, porque no se dispone todavía de herramientas generales para la visualización gráfica de la estructura del HL, ni de un tiempo razonable de aplicación y uso. Sin embargo, de los resultados y del proceso de construcción se han extraído algunas conclusiones que sintetizamos a continuación:

1. la riqueza relacional del tesauro depende del tipo de estructuras-t que se utilizan como fuente de datos para el tesauro. En el primer caso práctico, que trataba la especialización del tesauro ETB, las estructuras-t sólo contenían un tipo de relación, TG/TE, y, en consecuencia, se generará un tesauro de estructura simple de tipo clasificación o taxonomía, y con una capacidad limitada para ayudar al usuario a encontrar los recursos didácticos, puesto que no dispone de términos

equivalentes o términos cercanos semánticamente. Sin embargo, en el segundo caso práctico, el resultado ha sido un tesauro del repositorio CHASQUI rico en términos, categorías y relaciones, complejo en estructura con facetas, jerarquías de categorías y redes de términos TG/TE y TR, aunque apenas se utiliza la relación de equivalencia, probablemente debido a la especificidad del dominio (arqueología precolombina). La diferencia entre estos dos resultados está en que los autores de las estructuras-t del repositorio CHASQUI utilizaron un lenguaje de clasificación de los objetos virtuales más complejo que el que se recomienda para clasificar los objetos del repositorio AGREGA. En un punto medio, se encuentra el tesauro del Glosario e-Derecho obtenido en el tercer caso práctico; este tesauro contiene mayoritariamente redes de términos TR, que estructuran el tesauro en lo que podemos considerar 38 familias semánticas;

2. la corrección y coherencia del contenido depende, también, de la corrección de las estructuras-t y de la uniformidad con la que los docentes usan el lenguaje de especialidad para crear dichas estructuras. Si las estructuras-t contienen errores, éstos aparecen en el tesauro. Sin embargo, la estrategia de considerar sólo los términos y estructuras más frecuentes, minimiza el impacto, tanto más cuanto mayor sea el número de estructuras-t insertadas. Además, la fase de revisión manual limpia el tesauro de errores e, incluso, permite detectar y corregir estos errores en las clasificaciones de procedencia que, en otro caso, serían difíciles de detectar. La fase de revisión y corrección, por lo tanto, corrige no sólo el tesauro sino, también, los recursos o contenidos asociados;
3. el método HL no tiene previsto un sistema de control del vocabulario, porque entendemos que es el equipo de personas que crea las estructuras-t el que debe definir y mantener las reglas de control en la revisión y las relaciones de equivalencia entre términos que necesiten. Sin embargo, la falta de control puede generar problemas de dispersión, ya que las variantes léxicas o variantes ortográficas de un término que aparecen en las estructuras-t, se insertan en el tesauro, pero sin relacionar con sus formas preferidas o con sus formas alternativas. La experiencia de CHASQUI indica que: (i) una forma de controlar esta dispersión es introduciendo las correcciones durante la fase de revisión y adecuación o durante el uso del tesauro, (ii) las reglas de control surgen con el uso más frecuente de unos términos, términos preferidos, frente a otros no preferidos, y (iii) no es excesivamente costoso realizar estas correcciones

durante el uso del tesauro y se obtienen resultados buenos, y las tasas de dispersión son muy bajas; y

4. los tesauros obtenidos con el método HL tienen una estructura conceptual del dominio cercana a la de los profesores, puesto que se construyen con las estructuras-t que han creado. La experiencia de uso de estos tesauros, aunque escasa, nos ha mostrado algunas de las posibles líneas de explotación:
   a) facilitar la exploración del dominio para buscar y seleccionar los contenidos didácticos, gracias a que el lenguaje y la estructura conceptual es conocida por los usuarios y autores de las fuentes del tesauro;
   b) ayudar al experto a corregir y refinar la estructura terminológica del tesauro que representa la estructura conceptual del dominio de conocimiento al que pertenecen las estructuras-t;
   c) ayudar al experto a localizar y reutilizar los recursos o materiales didácticos y de investigación asociados al tesauro y a documentar nuevos recursos o materiales reutilizando o actualizando el tesauro;
   d) proporcionar un marco conceptual coherente y expresado con una terminología familiar del ámbito del tesauro, los recursos y materiales académicos; y
   e) se puede utilizar para estudiar la naturaleza del lenguaje de especialidad usado para describir las nociones de un dominio de conocimiento.

9.2. Conclusiones finales

El modelo y método HL es una herramienta de apoyo a la construcción de tesauros que, lejos de constituir una aproximación separada o independiente de los métodos de construcción de tesauros elaborados y experimentados durante décadas en el campo de la Biblioteconomía y Documentación, la Lingüística y la Informática, especializa estas aproximaciones en un nuevo entorno de aplicación, el entorno académico del *e-learning,* con el fin de facilitar a sus usuarios la construcción y gestión de una nueva forma de tesauro que hemos denominado tesauro académico de explotación.

El tesauro académico de explotación recoge y expresa, con el lenguaje propio de una comunidad científico-académica, los términos y conceptos relacionados con un dominio concreto. Proceden de vocabularios o de estructuras terminológicas en semántica libre, creados empíricamente por los profesores para expresar y organizar las ideas propias

sobre los contenidos y las colecciones de recursos que desarrollan individual o colectivamente.

El modelo y método HL se ha desarrollado, fundamentalmente, para que los equipos docentes puedan crear, de la forma más sencilla posible, estos recursos lingüísticos. Aportan cinco ventajas que enumeramos a continuación:

1. El modelo HL es una herramienta para conceptualizar la naturaleza, el funcionamiento y la evolución de los tesauros. A diferencia de otros modelos utilizados para la representación de tesauros, el modelo HL proporciona un formalismo de representación uniforme y, al mismo tiempo, adaptado a las características especiales del conocimiento léxico, altamente interrelacionado y en continua evolución. En concreto aporta:
    1.1. *un esquema de datos que es independiente* del tipo de tesauro, del tipo de contenido y del método de construcción. Puede considerarse un metamodelo, que proporciona un conjunto de constructores y reglas para crear modelos específicos de tesauros, capaces de i) ajustarse al lenguaje y al dominio de alcance y ii) reajustarse con los cambios que van surgiendo en el dominio y en el lenguaje;
    1.2. *un esquema de datos que es capaz de integrar, de forma coherente, estructuras parciales de términos relacionados semánticamente*, las estructuras-t, que están dispersas, inicialmente, en diversas fuentes de datos. Esta capacidad permite la construcción inductiva y continuada de los tesauros, unificando los procesos de mantenimiento que son costosos en el proceso de construcción; en este sentido, el modelo HL permite abandonar el concepto de construcción de un sistema léxico como sistema acabado y pasar a considerarlo como un sistema en permanente cambio, del que puede conocerse, en cada instante, su 'estado';
    1.3. *una semántica uniforme* para interpretar los signos del esquema de datos, basada en la noción del valor del significado de los términos, dependiente de la posición diferencial en el HL, y en la composición de significados para interpretar las categorías y el tesauro globalmente;
    1.4. *una representación visual, gráfica*, basada en el modelo lingüístico estructuralista de sistema de signos y en el modelo matemático de conjuntos y grafos, el higraph, que contribuye a entender la estructura y funcionamiento de los signos en el complejo sistema lingüístico que forman los tesauros de signos asociados por relaciones diferentes. Estas representaciones gráficas permiten,

además, descubrir nuevas estructuras contenidas en los tesauros, que no estaban explícitamente definidas en el contenido del tesauro y que con otros modelos no hubiera sido posible obtener;

1.5. *un modelo lingüístico formal* que abre nuevas posibilidades al estudio teórico de los sistemas lingüísticos aprovechando la teoría de conjuntos y grafos; y

1.6. *un modelo informático genérico* que no impone una notación ni un modelo de implementación para construir los tesauros como sistemas informáticos; de hecho, el modelo HL apoya nuevas formas de construcción de tesauros, especialmente de tipo inductivo.

2. El método HL es una herramienta para ayudar a construir los tesauros en dominios de aplicación con fuentes de estructuras-t.
3. El método HL es que es un método inductivo que simplifica los métodos tradicionales de construcción de tesauros aprovechando las fuentes de estructuras-t y el modelo general HL de representación léxica.
4. El método HL aporta una nueva forma de crear y mantener tesauros diferente de la concepción tradicional de crear los tesauros de forma reglada, sujeta a un diseño conceptual, elaborado de antemano o con posterioridad a la recogida de términos, a unas reglas de normalización, de presentación, de revisión y mantenimiento elaboradas por una autoridad responsable de la edición del tesauro. En esta aproximación el tesauro no se diseña, va surgiendo según se construye o actualiza, las reglas de normalización son las que regulan el lenguaje específico con el que los usuarios expresan sus ideas sobre el dominio, la presentación no tiene por qué estar predefinida pueden crearse dinámicamente, la revisión depende de los usuarios, y puede realizarse tanto en la construcción como durante el uso del tesauro, y finalmente, el mantenimiento se puede integrar en el proceso de creación de nuevos objetos del ámbito del tesauro.
5. El método HL, al contrario de otros métodos, está orientado a construir tesauros que reflejen el modo subjetivo en que los autores han organizado y descrito el ámbito del tesauro. Esta subjetividad se considera un inconveniente en la teoría y práctica general de construcción de tesauros porque afecta negativamente a la interoperabilidad y reusabilidad del contenido; sin embargo, en el ámbito académico proporciona tesauros más eficaces porque son más precisos y comprensibles para sus creadores y usuarios.

9.3. Líneas de trabajo futuro

"En general, puede afirmarse que no hay cuestiones agotadas,
sino hombres agotados en las cuestiones" (Ramón y Cajal, 1941)

Durante la elaboración de este trabajo han surgido innumerables cuestiones que no hemos podido resolver y que consideramos que merecen investigarse porque podrían dar lugar a resultados de interés teórico o práctico. De todas ellas vamos a destacar las cinco que nos parecen más interesantes por su relación más directa con esta tesis e incluso porque constituyen líneas en las que ya hemos comenzado a trabajar:

1. El modelo y método HL constituye el fundamento para la construcción de aplicaciones de carácter general de creación y gestión de tesauros. La aplicación práctica de las ideas desarrolladas en este trabajo sólo será posible si los profesores, investigadores y estudiantes disponen de una aplicación informática que, de forma transparente, implemente los tesauros académicos de explotación. Desgraciadamente, esta es una cuestión que no hemos podido abordar porque la magnitud de la tarea desborda las posibilidades de un investigador y de un trabajo de tesis. Este desarrollo, sin embargo, completaría el trabajo de investigación aquí presentado y facilitaría nuevas líneas de investigación aplicada, como las que siguen a este punto. Actualmente, esta línea forma parte de los objetivos de investigación de nuestro grupo, que tienen previsto la definición y construcción de un prototipo en el marco de un futuro proyecto de investigación financiado.
2. El método HL podría utilizarse en combinación con otros métodos de construcción de tesauros, para mejorar los resultados en aspectos concretos como por ejemplo, la interoperabilidad, o el rendimiento en ámbitos muy extensos como la Web que contienen tipos de estructuras-t muy diversas. En el primer caso, se podrían utilizar las técnicas de diseño de tesauros tradicionales para refinar los esquemas, inductivos, obtenidos con el método HL; en el segundo caso, podrían utilizarse técnicas de extracción automática de patrones de estructuras-t o técnicas de agrupamiento para crear las series semánticas sobre las que luego puede operar el método HL.
3. Otra línea de investigación, recientemente iniciada en el marco de un proyecto de innovación interdisciplinar, es el estudio del uso académico de los tesauros

académicos de explotación para apoyar al estudiante en el aprendizaje de los términos y nociones de una disciplina o área de conocimiento o para apoyar al experto en la localización de contenidos o recursos digitalizados que se refieran a un concepto; esta línea de trabajo no es nueva, pero todavía se desconoce cuál es su rentabilidad didáctica, la(s) forma(s) de uso en entornos *e-learning*, o cuáles son modelos los cognitivos que facilita al usuario el uso del tesauro y el aprendizaje de sus términos y nociones.

4. La aplicación del modelo y método HL para sistematizar obras poco estructuradas es, también, una cuestión directamente relacionada con este trabajo de tesis. Esta sistematización proporcionaría un mecanismo de acceso eficaz a contenidos que de forma creciente se están generando con la participación colaborativa y libre de los usuarios de los nuevos entornos de trabajo e interacción social creados con las nuevas tecnologías de la Web 2.0.
5. Finalmente, otra de las posibilidades, de carácter más teórico, abierta con este trabajo es la aplicación del modelo HL para el estudio de las estructuras léxicas que contienen los tesauros, aprovechando la perspectiva matemático-visual que aporta este modelo a la representación de la naturaleza y funcionamiento de estos subsistemas lingüísticos.

Bibliografía

AAT, 1994: Art and Architecture Thesaurus (5 vols.) 1994, 2nd ed., Oxford University, New York Disponible en: http://www.getty.edu/research/conducting_research/vocabularies/aat/#sample.

Abiteboul, S., Buneman, P. y Suciu, D. 2000, "Data on the web. From relations to semistructured data and XML", Morgan Kaufmann Publishers, San Francisco, California.

Aguirre, S., Quemada, J. y Salvachua, J. 2004, "Mediadores e interoperabilidad en Elearning", *Actas V Encuentro Internacional sobre Educación, Capacitación profesional y Tecnologías de la Información,* Barcelona Disponible en: http://jungla.dit.upm.es/saguirre/publications/virtualEduca2004.pdf.

Aitchison, J. y Clarke, S.D. 2004, "The thesaurus: A historical viewpoint, with a look to the future", *Cataloging y Classification Quarterly,* vol. 37, no. 3/4, pp. 5-21 Disponible en: http://www.informaworld.com/smpp/content~db=all~content=a903588884.

Aitchison, J., Gilchrist, A. y Bawden, D. 2000, "Thesaurus construction and use: a practical manual", Europa Publications

Allen, J. 1995, "Natural language understanding", Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA (2nd edition), (1st. edition 1988)

Alonso, J. 2001, "Recursos electrónicos: TESUAROS, CLASIFICACIONES, ENCABEZAMIENTOS", publicado en el foro Infodoc, Universidad de Salamanca, Facultad de Traducción y Documentación, Biblioteca, Disponible en: http://listas.bcl.jcyl.es:81/read/messages?id=768

Anderson, T. y Elloumi, F. 2004, "Theory and Practice of Online Learning", Athabasca University, Disponible en: http://cde.athabascau.ca/online_book.

ANECA 02_080314, 2008, *Programa Academia. Preguntas frecuentes sobre el modelo de evaluación 02_080314* Disponible en: http://www.aneca.es/active/docs/academia_faq02_080314.pdf.

ANSI/NISO Z39.19, 2005, *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, NISO Press, Bethesda, Maryland, U.S.A. Disponible en: http://www.niso.org/kst/reports/standards?step=2ygid=Noneyproject_key=7cc9b583cb5a62e8c15d3099e0bb46bbae9cf38a.

Antelman, K., Lynema, E. y Pace, A.K. 2006, "Toward a 21st Century Library Catalog", *Information Technology and Libraries,* vol. 25, no. 3 Disponible en: http://www.lib.ncsu.edu/staff/kaantelm/antelman_lynema_pace.pdf; http://www.ala.org/ala/mgrps/divs/lita/ital/252006/number3september/2503sept.cfm; http://www.ala.org/ala/mgrps/divs/lita/ital/252006/number3september/antelman.pdf.

Arano, S. y Codina, L. 2004, "La estructura conceptual de los tesauros en el entorno digital: ¿nuevas esperanzas para viejos problemas?", *9as. Jornadas Catalanas de información y documentación*Barcelona Disponible en: http://www.lluiscodina.com/ontotesauros.doc.

Area, M., San-Nicolás, B. y Fariña, E. 2008, *Evaluación del Campus Virtual de la Universidad de La Laguna: Análisis de las Aulas virtuales. Periodo 2005-07*, Universidad de La Laguna Disponible en: http://webpages.ull.es/users/manarea/informeudv.pdf.

ARIADNE Asociación Europea para compartir y reutilizar conocimiento. Disponible en: http://www.ariadne-eu.org/.

Ashman, H. y Simpson, R.M. 1999, "Computing surveys' electronic symposium on hypertext and hypermedia", *ACM Computing Surveys,* vol. 31, no. 4, pp. 325-334.

Austin, D. 1976, "Precis: a manual of concept analysis and subject indexing", *Journal of Librarianship and Information Science,* vol. 8, no. 3, pp. 210-212.

Bachman, C.W. 1972, "The evolution of storage structures", *Communications of the ACM,* vol. 15, no. 7, pp. 628-634 Disponible en: http://doi.acm.org/10.1145/361454.361495; http://delivery.acm.org/10.1145/370000/361495/p628-bachman.pdf?key1=361495ykey2=1217292521ycoll=GUIDEydl=GUIDEyCFID=51660118yCFTOKEN=90429366.

Baeza-Yates, R. y Ribeiro-Neto, B. 1999, "Modern Information Retrieval", Addison-Wesley, Wokingham, UK Disponible en: http://sunsite.dcc.uchile.cl/irbook/.

Balasubramanian, V. 1995, "State of the art review of hypermedia: issues and applications", Disponible en: http://www.e-papyrus.com/hypertext_review/index.html.

Balkan, L., Miller, K., Austin, B., Etheridge, A., Garcia Bernabé, M. y Miller, P. 2002, "ELSST: a broad-based Multilingual Thesaurus for the Social Sciences", *Third International Conference on Language Resources and Evaluation*, pp. 1873 Disponible en: https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2002/LREC/pdf/3.pdf.

Ballew, R., Duncan, T. y Blasingame, M. 1999, *Relational Data Structures for Implementing Thesauri*, Disponible en: http://www.mip.berkeley.edu/mip/related/thesaurus/thesaurus.pdf.

Banyard, P. y Underwood, J. 2008, "Understanding the learning space", *eLearning Papers,* vol. 9 Disponible en: http://www.elearningeuropa.info/files/media/media15970.pdf. Último acceso: julio 2008.

Barker, P. 2005, "What is IEEE Learning Object Metadata / IMS Learning Resource Metadata?", *CETIS guides to the IEEE Standard for Learning Object Metadata (LOM),* Disponible en: http://metadata.cetis.ac.uk/guides/WhatIsLOM.pdf.

Bechhofer, S., Harmelen, F.v., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P.F. y Stein, L.A. 2004, W3C Recommendation 10 February 2004, *OWL Web Ontology Language. Semantics and Abstract Syntax* Disponible en: http://www.w3.org/TR/owl-semantics/.

Beckett, D. W3C Recommendation 10 February 20042004, *RDF/XML Syntax Specification* Disponible en: http://www.w3.org/TR/rdf-syntax-grammar/.

Bennett, P. 2002, *Introduction to text categorization* Disponible en: http://www.softlab.ece.ntua.gr/facilities/public/AD/Text%20Categorization/Introduction%20to%20Text%20Categorization.ppt#256. Último acceso: 10/10/2007.

Berners-Lee, T., Hendler, J. y Lassila, O. 2001, "The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities", *Scientific American,* Disponible en: http://www.sciam.com/article.cfm?id=the-semantic-web.

Berrocal Román, F., Blanco Villalobos, J.J., Bouza Álvarez, M.T., Campo Moreno, S., Campo Vecino, J.d., Canet Parrilla, J.B., Casares del Río, José Miguel, Fuertes Royo, C., Galisteo del Valle, A., Gértrudix Barrio, M., Gómez Crespo, M.Á., González Prado, A., López, M., López Blanco, F., López de la Riva, América, Moreno Santos, J.A., Paredes Maña, María del Mar, Peña Sánchez, N., Porras Guardo, A.d., Retortillo Franco, F., Rodríguez San Felipe, María Jesús, Rojo García, E., Romero Montero, A., Sánchez Pedraza, J. y Vivancos Martí, J. 2008, *ANEXO X TAXONOMÍA DISCIPLINA TESAURO ETB MEC-CCAA V.1.0. 14-2-2008.* Disponible en: http://www.educa.madrid.org/cms_tools/files/799be97b-d80f-45b4-bda9-54bc1bb8d74c/a10_taxonomia_disciplina_etb_mec-ccaa_v1.pdf.

Bertino, E., Catania, B. y Zarri, G.P. 2001, "Intelligent Database System", Addison Wesley Longman Publishing Co., Boston, MA, USA.

Bertino, E. y Martino, L. 1995, "Sistemas de bases de datos orientadas a objetos", Addison-Wesley Iberoamericana.

Betancort, N. y Chozas, L. 2004, *Tesauros, mapas conceptuales y topic maps*. Disponible en: http://es.geocities.com/naolig/tesauros-mapas-conceptuales-topic-maps.htm

Bézivin, J. 2005, "On the Unification Power of Models", *Journal on Software and System Modeling,* vol. 4, no. 2, pp. 171-188 Disponible en: http://www.sciences.univ-nantes.fr/lina/atl/www/papers/OnTheUnificationPowerOfModels.pdf http://www.springerlink.com/content/xn50242535640k10/fulltext.pdf.

Bézivin, J. 2005, "On the Unification Power of Models", *Software and System Modeling,* vol. 4, no. 2, pp. 171-188 Disponible en: http://www.sciences.univ-nantes.fr/lina/atl/www/papers/OnTheUnificationPowerOfModels.pdf.

Bird, S., Hammond, M., Amarillas, M., Jeffcoat, M., Harley, H., Miyashita, M., Moll, L., Willie, M.A. y Zepeda, O. 2002, "Web-based Dictionaries for Languages of the South-west USA", *Literary and Linguistic Computing,* vol. 17, no. 4, pp. 427-238 Disponible en: http://llc.oxfordjournals.org/cgi/reprint/17/4/427.

Blustein, J. y Noor, M. 2004, "Personal glossaries on the WWW: an explanatory study", *Proceedings of the 2004 ACM Symposium on Document Engineering*, ed. J. Vion-Dury, ACM Press, New York, NY, USA, pp. 54 Disponible en: http://doi.acm.org/10.1145/1030397.1030409.

Borst, W.N. and Akkermans, J.M. and Top, J.L. (1997) Engineering Ontologies. International journal of human-computer studies, 46 (2-3). pp. 365-406. Disponible en: http://purl.org/utwente/18019

Bougarev, B.K. 1996, "Building a Lexicon: The Contribution of Computers", *International Journal of Lexicography,* vol. 4, no. 3, pp. 227-260.

Brachman, R.J. 1983, "What IS-A is and isn't: An analysis of taxonomic links in semantic networks", *IEEE Computer,* vol. 16, no. 10, pp. 30-36.

Brachman, R.J. y Levesque, H.J. (1985), "Readings in Knowledge Representation", Morgan Kauffmann, Los Altos, CA, USA.

Britain, S. y Liber, O. 2004, "A Framework for the Pedagogical Evaluation of eLearning Environments", *JISC-commissioned report,* Disponible en: http://www.cetis.ac.uk/members/pedagogy/files/4thMeet_framework/VLEfullReport

Brownson, H. 1957, "Proceedings of the International Study Conference on Classification for Information Retrieval", *Proceedings of the International Study Conference on Classification for Information Retrieval*, pp. 99.

Buckland, M., Chen, A., Chen, H., Kim, Y., Lam, B., Larson, R., Norgard, B. y Purat, J. 1999, "Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies", *D-Lib*

Magazine, vol. 5, no. 1 Disponible en: http://www.dlib.org/dlib/january99/buckland/01buckland.html.

Bush, V. 1945, "As We May Think", *The Atlantic,* Disponible en: http://www.theatlantic.com/doc/194507/bush.

Byrd, R.J., Calzolari, N., Chodorow, M.S., Klavans, J.L., Neff, M.S. y Rizk, O.A. 1987, "Tools and methods for computational lexicology", *American Journal of Computational Linguistics,* vol. 13, no. 3/4, pp. 219-240 Disponible en: http://acl.ldc.upenn.edu/J/J87/J87-3003.pdf.

Cáceres, J. 2007, "Estudio exploratorio de defectos en registros de meta-datos IEEE LOM de objetos de aprendizaje" En *Actas IV Simposio Pluridisciplinar sobre Diseño, Evaluación y Desarrollo de Contenidos Educativos Reutilizables (SPDECE*). Universidad del País Vasco, Disponible en: http://spdece07.ehu.es/actas/Caceres.pdf

Calzolari, N. 1991, "Lexical Databases and Textual Corpora: perspectives of integration for a lexical knowledge base" in *Lexical Acquisition. Exploiting on-line resources to build a lexicon*, ed. U. Zernik, Lawrence Erlbaum Associates Publishers, London, UK.

Calzolari, N. 1994, "Issues for Lexicon Building", en A. Zampolli, N. Calzolari & M. Palmer (eds.) *Current Issues in Computational Linguistics: In Honour of Don Walker.* Vol. IX-X, Pisa: Giardini Editori e Stampatori.

Carabantes, D., Carrasco, A. y Alves, J.D. 2005, "La innovación a través de entornos virtuales de enseñanza y aprendizaje", *Revista Iberoamericana de Educación a Distancia,* vol. 8, no. 1 y 2 Disponible en: http://www.utpl.edu.ec/ried/images/pdfs/vol8-1-2/innovacion_entornos.pdf.

Carbonell, J., Mitamura T. y Nyberg, E. H. 1992, "The Kant Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics,...)", en *Proceedings of the 4th Conference on Theoretical and Methodological Issues in Machine Translation (TMI '92),* 225-235.Disponible en: http://www.cs.cmu.edu/~jgc/publication/The_KANT_Prespective_A_Critique_ICTMI_1992.pdf

Casares, J. 1942, "Diccionario ideológico de la Lengua Española", Ed. Gustavo Pili, Barcelona.

CEN CWA 14871, 2003, *Controlled Vocabularies for Learning Objects Metadata: Typology, impact analysis, guidelines and a web based Vocabulary Registry* Disponible en: http://www.cen-isss-wslt.din.de/sixcms_upload/media/3050/CWA14871.pdf.

CEN CWA 15453, 2005, *Harmonisation of vocabularies for eLearning* Disponible en: ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/WS-LT/cwa15453-00-2005-Nov.pdf.

Centelles, M. 2005, "Taxonomías para la categorización y la organización de la información en los sitios web", *Hypertext.net,* no. 3 Disponible en: http://www.hipertext.net.

CERES thesaurus Disponible en: http://ceres.ca.gov/thesaurus/Overview.html.

Chen, P.P.S. 1976, "The Entity-Relationship Model: Toward a Unified View of Data", *ACM Transactions on Database Systems,* vol. 1, no. 1, pp. 9-36 Disponible en: http://citeseer.ist.psu.edu/519283.html.

Chen, Z., Liu, S., Wenyin, L., Pu, G., and Ma, W. 2003, "Building a web thesaurus from web link structure". In *Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval* (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM, New York, NY, 48-55. DOI= http://doi.acm.org/10.1145/860435.860447, http://portal.acm.org/citation.cfm?id=860447

CINDOC (Centro de Información y Documentación Científica) Tesauros Disponible en: http://thes.cindoc.csic.es/index_esp.html.

Clave, 2002, "Diccionario de uso del español actual", 2002, S. M., Madrid.

Codd, E.F. 1979, "Extending the Database Relational Model to Capture More Meaning", *ACM Transactions on Database Systems,* vol. 4, no. 4, pp. 397-434.

Codd, E.F. 1970, "Model of Data for Large Shared Data Banks", *Communications of the ACM,* vol. 13, no. 6, pp. 377-387 Disponible en: http://doi.acm.org/10.1145/362384.362685; http://delivery.acm.org/10.1145/370000/362685/p377-codd.pdf?key1=362685ykey2=5127292521ycoll=GUIDEydl=GUIDEyCFID=51660118yCFTOKEN=90429366.

Conklin, J. 1987, "Hypertext: An introduction and survey", *Computer,* vol. 20, no. 9, pp. 17-41.

Conole, G. y Fill, K. 2005, "A learning design toolkit to create pedagogically effective learning activities", *Journal of Interactive Media in Education,* vol. Portable Learning Special Issue Disponible en: http://www-jime.open.ac.uk/2005/08/conole-2005-08.pdf.

COSATI, 1967, *Guidelines for the development of information retrieval thesauri*; Sub-Panel on Classification and Indexing, Panel on Operational Techniques and Systems,

Committee on Scientific and Technical Information COSATI Washington, D.C. Disponible en: http://catalogue.nla.gov.au/Record/1537468

Cross, P., Brickley, D. y Koch, T. 2000, 06-Jun-00-last update, *Conceptual relationships for encoding thesauri, classification systems and organised metadata collections and a proposal for encoding a core set of thesaurus relationships using an RDF Schema*. Disponible en: http://www.desire.org/results/discovery/rdfthesschema.html.

Crouch, C. J. 1988. "A cluster-based approach to thesaurus construction". In *Proceedings of the 11th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Grenoble, France). Y. Chiaramella, Ed. SIGIR '88. ACM, New York, NY, 309-320. DOI= http://doi.acm.org/10.1145/62437.62467

Crouch, C. J. y Yang, B. 1992, "Experiments in automatic statistical thesaurus construction". In *Proceedings of the 15th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Copenhagen, Denmark, June 21 - 24, 1992). N. Belkin, P. Ingwersen, and A. M. Pejtersen, Eds. SIGIR '92. ACM, New York, NY, 77-88. DOI= http://doi.acm.org/10.1145/133160.133180 Disponible en: http://portal.acm.org/citation.cfm?id=133180#

Cruse, D.A. 2002, "Hyponymy and its varieties" in *The semantic relationships: an interdisciplinary perspective*, eds. R. Green, C.A. Bean y S.H. Myaeng, Kluwer Academic Publishers, Boston, pp. 35-50.

Cuesta Morales, P., Maña López, M.J. y Cuervo Martínez, C. 2000, "Aplicación de Técnicas de Recuperación de Información a un Glosario de Términos de Internet Desarrollado Utilizando Tecnologías JSP", *Actas de JBIDI, Primeras Jornadas de Bibliotecas Digitales*, ed. N. Brisaboa, Universidad de Valladolid, Departamento de informática, Valladolid, pp. 129 Disponible en: http://trevinca.ei.uvigo.es/~pcuesta/publicaciones/glosario.pdf.

DADS 2007, "Dictionary of Algorithms and Data Structure – DADS", National Institue of Standars and Technology - NIST 2007, Disponible en: http://www.itl.nist.gov/div897/sqg/dads/.

Dalmau, M., Floyd, R., Jiao, D. y Riley, J. 2005, "Integrating thesaurus relationships into search and browse in an online photograph collection", *Library Hi Tech,* vol. 23, no. 3, pp. 425-452 Disponible en: http://www.emeraldinsight.com/10.1108/07378830510621829 http://www.emeraldinsight.com/Insight/viewPDF.jsp?contentType=ArticleyFilename=html/Output/Published/EmeraldFullTextArticle/Pdf/2380230311.pdf.

D'Antoni, S. (ed.), 2003, "The Virtual University Models and messages Lessons: from case studies", UNESCO, Disponible en: http://www.unesco.org/iiep/virtualuniversity/home.php.

DCMI *The Dublin Core Metadata Registry* Disponible en: http://dublincore.org/dcregistry/.

Dekkers, M. y Feria, L. (eds.), 2006, *Metada for knowledge and learning: DC-2006, proceedings of the International Conference on Dublin Core and Metada Applications*, University of Colima, Department of Information Technologies, Mexico Disponible en: http://dcpapers.dublincore.org/ojs/pubs/issue/view/29.

Dichev, C. y Dicheva, D. 2006, "View-Based Semantic Search and Browsing", *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*IEEE Computer Society, Washington, DC, USA Disponible en: http://delivery.acm.org/10.1145/1250000/1249148/274700919.pdf?key1=1249148ykey2=7012952521ycoll=GUIDEydl=GUIDEyCFID=52238520yCFTOKEN=35770361.

Dondi, C. 2008, "La calidad de la experiencia de aprendizaje como factor discriminante en el desarrollo del potencial de las TIC en los sistemas educativos y formativos", *IV Jornada Campus Virtual UCM: experiencias en el Campus Virtual (Resultados)* Editorial Complutense, Madrid.

DRAE, 2001, "Diccionario de la lengua española" Real Academia Española (2001), Madrid , Espasa, 22.ª ed. 2001.

Dublin Core, 2008, *Dublin Core Metadata Element Set, Version 1.1* Disponible en: http://dublincore.org/documents/dces/.

Duffy, T. y Cunningham, D. 1996, "Implications for the design and delivery of instruction. Handbook of research for educational telecomunications and technology", McMillan, New York.

Duncan, E.B. 1990, "A concept-map thesaurus as a knowledge-based Hypertext interface to a bibliographic database", *Informatics 10: prospects for intelligent retrieval*Aslib, London, pp. 43.

ELRA 2003, *European Language Resources Association*. Disponible en: http://www.elra.info/

Elsmari, R. y Navathe, S.B. 1997, "Sistemas de bases de datos. Conceptos fundamentales", Addison-Wesley Iberoamericana.

Epper, R.M. y Garn, M. 2004, "The Virtual University in America: Lessons from Research and Experience", *Educase Centre For Applied Research (ECAR) Research Bulletin,* Disponible en: http://net.educause.edu/ir/library/pdf/ERB0402.pdf.

ERIC Thesaurus, Thesaurus of ERIC Descriptors. Educational Resource Information Center (ERIC) Processing y Reference Facility. Disponible en: www.eric.ed.gov/thesaurus.

ETB 2008, *European Tresaury Browser (European Schoolnet-ETB)*, European Commission, Bruselas Disponible en: http://etb.eun.org/etb/index.htm.

Feldman, R. y Sanger, J. 2007, "The text mining handbook: advanced approaches in analyzing unstructured data ", Cambridge University Press, New York, NY, USA.

Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D.L. y Patel-Schneider, P.F. 2001, "OIL: An Ontology Infrastructure for the Semantic Web", *IEEE Intellingent Systems,* vol. 16, no. 2, pp. 38-45 Disponible en: http://www2.computer.org/portal/web/csdl/magazines/intelligent;jsessionid=7092673ABF32DA8C4BD8908960A6E9F2#3; http://doi.ieeecomputersociety.org/10.1109/5254.920598; http://www.cs.vu.nl/~frankh/postscript/IEEE-IS01.pdf.

Fernández-Pampillón, A.M. 2005, "Herramientas de análisis textual: claves para el análisis de la "información" textual" in *Las nuevas profesiones de las lenguas*, eds. A. López Varela y A. Sanz, Liceus, Madrid, pp. 79-91.

Fernández-Pampillón, A.M., Fernández-Valmayor, A. y López-Alonso, C. 2003, "Representación y organización del conocimiento léxico: del modelo de datos hipertexto al modelo hiperred", *Tendencias de investigación en organización del conocimiento / Trends in knowledge organization research*, eds. C. Travieso Rodríguez y J.A. Frías Montoya.

Fernández-Pampillón, A.M. y Matesanz del Barrio, M. 2003, "Los diccionarios electrónicos: hacia un nuevo concepto de diccionario" in *Nuevos géneros discursivos: los textos electrónicos*, eds. C. López-Alonso y A. Séré, Estudios de Lingüística del Español (ELiEs), Disponible en: http://elies.rediris.es/elies24.

Fernández-Valmayor, A., Cristóbal, J., Navarro, A., Fernández-Pampillón, A.M., Peralta, M., Merino, J. y Roldán, Y. 2008, "El campus virtual en la universidad Complutense de Madrid", *PixelBit, revista de Medios y Comunicación,* vol. 32: Monográfico TIC y Universidad, pp. 55-65 Disponible en: http://www.sav.us.es/pixelbit/pixelbit/articulos/n32/4.pdf.

Fernández-Valmayor, A.; Fernández-Pampillón, A.M. y Merino, J. (eds.), 2007, *III Jornada Campus Virtual UCM: Innovación en el Campus Virtual metodologías y*

herramientas, Editorial Complutense, Madrid Disponible en: http://eprints.ucm.es/5835/.

Fernández-Valmayor, A.; Sanz, A.; y Merino, J. (eds.), 2008, *IV Jornada Campus Virtual UCM: experiencias en el Campus Virtual (Resultados)*, Editorial Complutense, Madrid Disponible en: http://eprints.ucm.es/7773/1/ACTAS_campusvirtual.pdf.

Fillmore, C.J. 1968, "The Case for Case" in *Universals in Linguistics Theory*, eds. E. Bach y R.T. Harms, Holt, Rinehart and Winston, New York, pp. 1-88.

Flores Doña, M.S. 2009, "Pedagogical innovation applied to Commercial law", E-prints Complutense, Madrid, Spain Disponible en: http://eprints.ucm.es/8815/; http://eprints.ucm.es/8815/1/EPrint_M%C2%AA_Sierra.pdf.

Flores Doña, M.S., Fernández-Pampillón, A.M., López Orozco, J.A. y Matesanz del Barrio, M. 2009, "El Glosario E-Derecho: un modelo empírico de información jurídica para la transmisión y comprensión del Derecho de Propiedad Intelectual en los campus virtuales universitarios", *Memorias de la Octava Conferencia Iberoamericana en Sistemas, Cibernética e Informática: CISCI 2009* Disponible en: http://eprints.ucm.es/8831/2/Congreso_glosarioEderecho_(3).pdf.

Fogarty, K. 2006, "System modeling and traceability applications of the higraph formalism". *Thesis, Department Systems Engineering, University of Maryland (College Park, Md.)* Disponible en: http://hdl.handle.net/1903/3560

Friesen, N. 2004, *Final Raport on the "International LOM Survey" SC 36_WG4_N0109 ISO/IEC JTC1/SC36/WG4N0109*, Canada Disponible en: http://mdlet.jtc1sc36.org/doc/SC36_WG4_N0109.pdf.

Fuller, B. 1962, "Education automation. Freeing the scholar to return to his studies", Southern Illinois University Press, Londres y Amsterdam.

Ganzmann, J. 1990, "Criteria for the evaluation of thesaurus software", *International Classification,* vol. 17, no. 3/4, pp. 148-157 Disponible en: http://www.willpowerinfo.co.uk/ganzmann.htm.

Garshol, L.M. 2004, *Metadata? Thesauri? Taxonomies? Topic Maps!* Disponible en: http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html.

GEM (GATEWAY to EDUCATIONAL MATERIALS). Disponible en: www.thegateway.org.

GEM (Gateway to Educational Materials) Controlled Vocabularies Disponible en: http://www.thegateway.org/about/documentation/gem-controlled-vocabularies/.

Génova, G., Lloréns, J., Fuentes, J.M., Morato, J. y Martínez, P. 2000, "Conceptual hierarchies in UML: comparing ISO 2788 standard with the UML metamodel", Disponible en: http://www.ie.inf.uc3m.es/ggenova/pub-ecoop2000.html.

Gibbon, D. 2000, "Computational Lexicography" in *Lexicon Development for Speech and Language Processing*, eds. F.v. Eynde y D. Gibbon, ELSNET, Kluwer Academic Publishers, pp. 1-42.

Gil Urdiciain, B. 1998a, "Evaluación semántica y estructural de tesauros", *Revista general de información y documentación,* vol. 8, no. 2, pp. 193-199 Disponible en: http://revistas.ucm.es/byd/11321873/articulos/RGID9898220193A.PDF.

Gil Urdiciain, B. 1998b, "Orígenes y evolución de los tesauros en España", *Revista general de información y documentación,* vol. 8, no. 1, pp. 63-110 Disponible en: http://revistas.ucm.es/byd/11321873/articulos/RGID9898120063A.PDF.

Gilchrist, A. 2003, "Thesauri, taxonomies and ontologies - an etymological note", *Journal of Documentation,* vol. 59, no. 1, pp. 7-18.

Gómez Hidalgo, J.M., Cortizo Pérez, J.C., Puertas Sanz, E. y Buenaga Rodríguez, M.d. 2004, "Experimentos en indexación conceptual para la categorización de texto", *Actas de la Conferencia Iberoamericana WWW/Internet 2004*, eds. J.M. Gutiérrez, J.J. Martínez y P. Isaías, Madrid Disponible en: http://www.ainetsolutions.com/jccp/papers/ciawi04a.pdf.

Goñi Menoyo, J.M. 1998, *Arquitectura para la representación del conocimiento léxico en sistemas de procesamiento del lenguaje natural*, Departamento de Ingeniería de Sistemas Telemáticos. Tesis doctoral presentada en la Escuela Técnica Superior de Ingeniero de Telecomunicaciones. Universidad Politécnica de Madrid Disponible en: http://oa.upm.es/91/01/Tesis.pdf.

Greenberg, J., Heidorn, B., Seiberling, S. y Weakley, A.S. 2005, "Growing vocabularies for plant identification and scientific learning", Universidad Carlos III de Madrid, Madrid, Spain, pp. 99 Disponible en: http://dcpapers.dublincore.org/ojs/pubs/issue/view/28; http://dcpapers.dublincore.org/ojs/pubs/article/view/808/804.

Grefenstette, G. 1994, "Explorations in Automatic Thesaurus Discovery", *Kluwer International Series in Engineering and Computer Science*Kluwer Academic Publishers, Norwell, MA, USA.

Griffiths, D., Blat, J., García, R. y Sayago, S. 2004, "La aportación de IMS Learning Design a la creación de recursos pedagógicos reutilizable", , eds. J.R. Hilera González, Gutiérrez de Mesa, José A., R. Vélez de Miguel y R. Martínez Borda, Universidad de Alcalá, Alcalá de Henares (Madrid).

Gruber, T.R. 1993, "A translation approach to portable ontology specification", *Knowledge adquisition,* vol. 5, no. 2, pp. 199-220 Disponible en: http://ksl-web.stanford.edu/knowledge-sharing/papers/README.html#ontolingua-intro.

Gruninger, M. y Lee, J., 2002, "Ontology applications and Design". *Communications of the ACM.* February 2002, vol. 45, No. 2 pp.39 Disponible en: http://portal.acm.org/citation.cfm?id=503124.503145

Guinea Bueno, M. 2004, "El proyecto Chasqui", *Campus Virtual UCM*, eds. A. Fernández-Valmayor, A.M. Fernández-Pampillón y J. Merino, Editorial Complutense, Madrid, pp. 228.

Guri-Rosenblit, S. 2005, "Eight Paradoxes in the Implementation Process of E-learning in Higher Education", *Higher Education Policy,* vol. 18, pp. 5-29 Disponible en: http://www.smkb.ac.il/privweb/chaim_tir/meds/eight.pdf.

Guthrie, L., Pustejovsky, J., Wilks, Y. y Slator, B.M. 1996, "The role of lexicons in natural language processing", *Communications of the ACM,* vol. 39, no. 1, pp. 63-72 Disponible en: http://doi.acm.org/10.1145/234173.234204; http://delivery.acm.org/10.1145/240000/234204/p63-guthrie.pdf?key1=234204ykey2=9066292521ycoll=GUIDEydl=GUIDEyCFID=51659098yCFTOKEN=72208960.

Haensch, G. 1997, "Los diccionarios del español en el umbral del siglo XXI", Ediciones Universidad de Salamanca, Salamanca.

Hall, B. 2007, *LMS and LCMS Demystified* Disponible en: http://www.brandon-hall.com/free_resources/lms_and_lcms.shtml.

Harel, D. 1988, "On visual formalisms", *Communications of the ACM,* vol. 31, no. 5, pp. 514-530 Disponible en: http://portal.acm.org/citation.cfm?id=42414.

Harmon, P. y Watson, M. 1998, "Understanding UML. The developer's guide", Morgan Kaufmann Publishers, Inc., San Francisco, California, US.

Hazewinkel, M. 1997, "Enriched thesauri and their uses in information retrieval and storage. Discussion paper", *Proceedings of the First DELOS Workshop: An Overview on Projects and Research Activities in Digital Library Related Fields*, ed. C. Thanos, ERCIM, pp. 27 Disponible en: http://www.ercim.org/publication/ws-proceedings/DELOS1/hazewinkel.pdf.

Heath, B., McArthur, D.J., McClelland, M.K. y Vetter, R.J. 2005, "Metadata Lessons from the iLuminia Digital Library", *Communications of the ACM,* vol. 48, no. 7.

HEFCE, 2005, HEFCE 2005/12, *HEFCE strategy for e-learning: policy development* Disponible en: http://www.hefce.ac.uk/pubs/hefce/2005/05_12/05_12.pdf.

Hepp, M. 2007, "Possible Ontologies. How Reality Constrains The Development of Relevant Ontologies", *IEEE Internet Computing,* vol. 11, no. 1, pp. 90-96 Disponible en: http://www.computer.org/portal/site/internet/.

Hernández, E. 2003, *Estándares y Especificaciones de E-learning: Ordenando el Desorden* Disponible en: http://www.uv.es/ticape/docs/eduardo.pdf.

Hernández, H. y Saiz, M. 2007, "Ontologías mixtas para la representación conceptual de objetos de aprendizaje", *Procesamiento del Lenguaje Natural,* , no. 38 Disponible en: http://www.sepln.org/revistaSEPLN/revista/38/11.pdf.

Hillmann, D.I., Phipps, J., Sutton, S.A. y Laundry, R. 2006, *A Metadata Registry from Vocabularies Up: the NSDL Registry Proyect* Disponible en: http://arxiv.org/ftp/cs/papers/0605/0605111.pdf.

Hirst, G. 2004, "Ontology and the Lexicon" in *Handbook on Ontologies*, eds. S. Staab y R. Studer, Springer, , pp. 209-230.

Holm, B.E. y Rasmussen, L.E. 1961, "Development of a technical thesaurus", *American Documentation,* vol. 12, no. 3, pp. 184-190.

Hortalá, M.T., Leach, J. y Rodríguez, M. 2001, "Matemática Discreta y Lógica Matemática", Editorial Complutense, .

Hovy, E.H., Philpot, A., Klavans, J.L., Germann, U., Davis, P.T. y Popper, S.D. 2003, "Extending Metadata Definitions by Automatically Extracting and Organizing Glossary Definitions", *Proceedings of NSF's national conference on Digital Government* Disponible en: http://www.isi.edu/natural-language/people/hovy/papers/03dgo-glossary-to-metadata.pdf.

Huggett, M. y Lanir, J. 2007, "Static reformulation: a user study of static hypertext for query-based reformulation", *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*ACM Press, New York, NY, USA, pp. 319.

Huynh, D., Mazzocchi, S. y Karger, D. 2005, "Piggi Bank: Experience the Semantic Web Inside Your Web Browser", *International Semantic Web Conference (ISWC) 2005* Disponible en: http://simile.mit.edu/papers/iswc05.pdf.

IEEE-LOM 1484.12.1, 2002, Working group 12. Computer Society/Learning Technology Standards Committee, *IEEE Standard for Learning Object Metadata,1484.12.1-2002* Disponible en: http://ltsc.ieee.org/wg12; http://ltsc.ieee.org/wg12/; http://www.imsproject.org/metadata/.

IMS 2007, *Abstract Framework-Glossary.* Disponible en: http://www.imsglobal.org/af/afv1p0/imsafglossaryv1p0.html.

IMS Digital Repositories, 2003, *Digital Repositories Specification v. 1.0 Final Specification* Disponible en: http://www.imsglobal.org/digitalrepositories/.

IMS Meta-data, 2004, *IMS Meta-data Best Practice Guide for IEEE1484. 12. 1-2002 Standard for Learning Object Metadata. Version 1.3. Final Specification* Disponible en: http://www.imsglobal.org/metadata/mdv1p3/imsmd_bestv1p3.html.

IMS VDEX Model, 2004, *Vocabulary Definition Exchange Information Model* Disponible en: http://www.imsglobal.org/vdex/.

IMS 2002, *IMS Global Learning Consortium Inc. Learning Resource Meta-data Specification Version 1.3* Disponible en: http://www.imsproject.org/metadata/index.html.

ISO 639, 1998, Codes for the representation of names of languages -- Part 2: Alpha-3 code. Disponible en: http://www.iso.org/iso/catalogue_detail?csnumber=4767

ISO 2788, 1986, *Guidelines for the establishment and development of monolingual thesauri*.

ISO 3166, 1999 *English country names and code elements* Disponible en: http://www.iso.org/iso/english_country_names_and_code_elements

ISO 5127, 2001, *Information and documentation -- Vocabulary* Disponible en: http://www.iso.org/iso/catalogue_detail.htm?csnumber=33636.

ISO/IEC 19501, 2005, *Information technology -- Open Distributed Processing -- Unified Modeling Language (UML) Version 1.4.2,* Disponible en: http://www.iso.org/iso/catalogue_detail.htm?csnumber=32620; también en: http://www.omg.org/cgi-bin/doc?formal/05-04-01

ISO/IEC 9075-14, 2008, *Information technology -- Database languages -- SQL -- Part 14: XML-Related Specifications (SQL/XML)* Disponible en: http://www.iso.org/. Se puede consultar la versión (libre) de Oracle: SQL Language Reference (2008) 11g Release 1 B28286 Mayo 2008 Disponible en:http://download.oracle.com/docs/cd/B28359_01/server.111/b28286.pdf

Jenkins, M., Browne, T. y Walker, R. 2005, "VLE Surveys: a longitudinal perspective between March 2001, March 2003 and March 2005 for Higher Education in the UK", Disponible en: http://www.ucisa.ac.uk/groups/tlig/vle/vle_survey_2005.pdf.

Joffe, D. y Schryver, G. 2004, "TshwaneLex A State-of-the-Art Dictionary Compilation Program", *Conference Proceedings of EURALEX 2004* Disponible en: http://tshwanedje.com/publications/euralex2004-TL.pdf.

Jones, S. 1993, "A thesaurus data model for an intelligent retrieval system", *Journal of Information Science,* vol. 9, no. 3, pp. 167-178.

Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J. y Walker, S. 1995, "Interactive thesaurus navigation: Intelligence rules OK?", *Journal of the American Society for Information Science,* vol. 46, no. 1, pp. 53-59.

Jovanovic, J., Gaševic, D., Verbert, K. y Duval, E. 2005, "ALOCOM Ontology. Ontology of Learning Object Content Structure", *Proceedings of the 12th International Conference on Artificial Intelligence in Education,* Amsterdam, The Netherlands.

Kerschberg, L. y Weishar, D.J. 2000, "Conceptual Models and Architectures for Advanced Information Systems", *Applied Intelligence,* vol. 13, no. 2, pp. 149-164 Disponible en: http://www.springerlink.com/content/gg50055223045722/fulltext.pdf; http://portal.acm.org/citation.cfm?id=590928.

Khan, B.H. 2005, "Learning Features in an Open, Flexible, and Distributed Environment", *AACE Journal,* vol. 13, no. 2, pp. 137-153 Disponible en: http://asianvu.com/digital-library/elearning/Learning_Features_in_Open_Learning_Badrul_Khan.pdf.

Kimoto, H. y Iwadera, T. 1990, "Construction of a Dinamyc Thesaurus and Its Use for Associated Information Retrieval", *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, ed. J. Vidick , ACM, New York, NY, USA, pp. 227.

Knight, C., Gaševic, D. y Richards, G. 2005, "Ontologies to integrate learning design and learning content", *Journal of Interactive Media in Education,* vol. 07, no. Advances in Learning Design: Special Issue Editorial (eds. Colin Trattersall, Rob Koper) Disponible en: http://www-jime.open.ac.uk/2005/07/knight-2005-07.pdf.

Koper, R. 2005, "An Introduction to Learning Design" in *Learning design: a Handbook on Modelling and Delivering Networked Education and Training*, eds. R. Koper y C. Tattersall, Springer, Heidelberg.

Laguens García, J.L. 2006, "Tesauros y lenguajes controlados en Internet", *Anales de Documentación,* vol. 9, pp. 105-121 Disponible en: http://revistas.um.es/analesdoc/article/viewFile/1391/1441; http://redalyc.uaemex.mx/redalyc/pdf/635/63500907.pdf.

Lamarca, M.J. 2007, *Hipertexto, el nuevo concepto de documento en la cultura de la imagen*, Facultad de Ciencias de la Información. Dpto. de Biblioteconomía y Documentación. Universidad Complutense de Madrid Disponible en: http://www.hipertexto.info.

Lancaster, F.W. 1986, "Vocabulary Control for information retrieval", Information Resource Press. 1992, 2ª edición.

Lancaster, F. W. y Warner, A. J. 1993 "Information Retrieval Today". 1st edition Information Resources Press.

LDC (Linguistic Data Consortium). Disponible en: http://www.ldc.upenn.edu/.

Lee, W. y Sugimoto, S. 2005, "Toward Core Subject vocabularies for Community-oriented Subject Gateways", *Vocabularies in Practice. DC-2005: Proceedings of the International Conference on Dublin Core and Metadata Applications*, ed. E. Méndez, Universidad Carlos III, Madrid, España.

Leech, G. 2005, "Adding Linguistic Annotation" in *Developing Linguistic Corpora: A Guide to Good Practice*, ed. M. Wayne, Oxbow Books, Oxford, pp. 17-29 Disponible en: http://ahds.ac.uk/linguistic-corpora.

Lehmann, A. y Martin-Berthet, F. 1998, "Introduction à la Lexicologie", Dunod, Paris, France.

Lewis, D.D. y Sparck Jones, K. 1996, "Natural Language Processing for Information Retrieval", *Communications of the ACM,* vol. 39, no. 1, pp. 92 Disponible en: http://www.informatik.uni-trier.de/~ley/db/journals/cacm/cacm39.html#LewisJ96; http://cacm.acm.org/magazines/1996/1; http://delivery.acm.org/10.1145/240000/234210/p92-lewis.pdf?key1=234210ykey2=8425292521ycoll=portalydl=ACMyip=147.96.41.83yCFID=51656280yCFTOKEN=27968092; http://portal.acm.org/citation.cfm?id=234173ycoll=portalydl=ACMytype=issueyidx=J79ypart=magazineyWantType=Magazinesytitle=Communications%20of%20the%20ACMyCFID=51656280yCFTOKEN=27968092.

Liu, L., Halper, M., Geller, J. y Perl, Y. 1999, "Controlled Vocabularies in OODBs: Modeling Issues and Implementation", *Distributed and Parallel Databases,* vol. 7, no. 1, pp. 37-65 Disponible en: http://www.springerlink.com/content/r51135r880633622/fulltext.pdf; http://portal.acm.org/citation.cfm?id=607191.

LOM-ES, 2008, LOM-ES v.1.0:2008, *GT9 / GT8 - SC 36/AENOR PERFIL DE APLICACIÓN LOM-ES1 V.1.0.* Disponible en: http://www.educa.madrid.org/cms_tools/files/ac98a893-c209-497a-a4f1-93791fb0a643/lom-es_v1.pdf.

López, M.A. y Moreira, J.A. 2007, *Presente y futuro de los tesauros como herramienta conceptual de precisión para la recuperación de la información*, Biblioteca Digital. Universitat Pompeu Fabra Disponible en: http://docdigital.upf.es/digital/aula2000/aula.htm.

López-Alonso, C., Matesanz, M. (Eds.) 2009, "Las plataformas de aprendizaje. Del mito a la realidad", Editorial Biblioteca Nueva. Madrid 2009.

López Alonso, C., Fernández-Pampillón, A., de Miguel, E., y Matesanz, M., 2009, "E-Ling: un paradigma del triángulo del conocimiento". En C. López Alonso y M. Matesanz del Barrio (eds.), Aprender en una plataforma. Del mito a la realidad. Madrid: Biblioteca Nueva.

López-Alonso, C., Fernández-Pampillón, A.M., Miguel, E.d. y Pita, G. 2008a, "Learning to research in a Virtual Learning Environment: a socio-constructivist mode", En Papadopoulos, G. A., Wojtkowski, W., Wojtkowski, W. G., Wrycza, S., & Zupancic, J. (eds) *Information Systems Development: Towards a Service Provision Society*, Springer-Verlag: New York., Disponible, también, en: http://eprints.ucm.es/8109/1/Microsoft_Word_-_paper118.doc.pdf.

López-Alonso, C., Miguel, E.d. y Fernández-Pampillón, A.M. 2008b, "Propuesta de integración de LAMS en el marco conceptual del espacio de aprendizaje socio-constructivista E-Ling", *2008 European LAMS Conference*.

López-Alonso, C. y Séré, A. 2005, "GALANET: una plataforma de enseñanza multimedia interactiva para la intercomprensión en lenguas románicas" in *Palabras, norma, discurso: en memoria de Fernando Lázaro Carreter*, eds. L. Santos Ríos, J. Borrego Nieto, J.F. García Santos, J.J. Gómez Asencio y Prieto de los Mozos, Emilio, Ediciones Universidad de Salamanca, Salamanca, pp. 695-710.

Lorentz, D. 11g Release 1 B282862008, *SQL Language Reference* Disponible en: http://download.oracle.com/docs/cd/B28359_01/server.111/b28286.pdf.

Lorite, A. 2004, "Depuración, traducción y adaptación del lenguaje de indización de la Biblioteca y Centro de Documentación (BDC) del CES de España", Disponible en: http://www.ces-galicia.org/jornadas/j2/6.doc.

Lyons, J. 1977, "Semantic", Cambridge University Press, Cambridge, Great Britain. 2ª edición (en español) 1989, "Semántica", Editorial Teide, Barcelona

Manning, C.D., Jansz, K. y Indurkhya, N. 2001, "Kirrkirr: Software for Browsing and Visual Exploration of a Structured Warlpiri Dictionary", *Literary and Linguistic Computing,* vol. 16, no. 2, pp. 135-151 Disponible en: http://llc.oxfordjournals.org/cgi/reprint/16/2/135.

Mao, W. y Chu, W.W. 2007, "The phrase-based vector space model for automatic retrieval of free-text medical documents", *Data y Knowldege Engineering,* vol. 61, no. 1, pp. 76-92.

Martínez de Sousa, J. 1995, "Diccionario de lexicografía práctica", Bibliograf, S.A., Barcelona.

Martínez, F.J. y García, J.C. 2006, *Tesauro de Redes de Ordenadores* Disponible en: http://www.um.es/gtiweb/fjmm/tesauro/index.html.

Martínez, F.J., Martínez, L. y Rodríguez, J.V. 1992, "Diseño Lógico-Conceptual de Tesauros", *Actas IV Jornadas Catalanas de Documentación*Barcelona, pp. 341 Disponible en: http://www.um.es/gtiweb/fjmm/disetesa.htm.

Marzal, M.Á., Colmenero, M.J., Calzada, J. y Cuevas, A. 2006, "Mapas conceptuales y presentación gráfica del tesauro: aplicación a las bibliotecas educativas", *Procceedings of the Second International Conference on Concept Mapping - Concept Maps: Theory, Methodology, Technology*, eds. A.J. Cañas y J.D. Novak, Disponible en: http://cmc.ihmc.us/cmc2006Papers/cmc2006-p70.pdf.

Matthews, B., Miles, A. y Wilson, M. 2003, *Modelling Thesuri for the Semantic Web. Workshop on Semantic Web and database 2003* Disponible en: http://www.w3c.rl.ac.uk/SWAD/papers/thesaurus/swdbpaper.html.

McCray, A.T. 2003, "An upper-level ontology for the biomedical domain", *Comparative and Functional Genomics, Comp Funct Genom* 2003; 4: 80–84, Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cfg.255. Disponible en: http://lhncbc.nlm.nih.gov/lhc/docs/published/2003/pub2003023.pdf

McGraw, K.L. 2001, "E-Learning Strategy Equals Infrastructure", *Learning Circuits,* Disponible en: http://www.learningcircuits.org/2001/jun2001/mcgraw.html http://www.astd.org/LC/2001/0601_mcgraw.htm.

McGuinness, D.L. 2003, "Ontologies Come of Age" in *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, eds. D. Fensel, J. Hendler, H. Lieberman y W. Wahlster, MIT Press, Disponible en: http://www-ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-(with-citation).htm.

Méndez, E. (ed.), 2005, *Vocabularies in Practice. DC-2005: Proceedings of the International Conference on Dunlin Core and Metadata Applications*, Universidad Carlos III, Madrid, España.

MESH, 2007*, Medical Subject Headings. National Library of Medicine*, Disponible en: http://www.nlm.nih.gov/mesh/2007/index.html.

Metaxides, A., Helgeson, W.B., Seth, R.E., Bryson, G.C., Coane , M.A., Dodd, G.G., Earnest , C.P., Engles, R.W., Harper, L.N., Hartley, P.A., Hopkin, D.J., Joyce, J.D., Knapp, S.C., Lucking, J.R., Muro, J.M., Persily, M.P., Ramm, M.A., Russell, J.F., Schubert, R.F., Sidlo, J.R., Smith, M.M. y Werner, G.T. 1971, Codasyl Technical Reports, April 19711971, *Data base task group report to the CODASYL programming language committee, April 1971*, ACM, New York.

Miguel, A. de, y Piattini, M. 1997, "Fundamentos y modelos de bases de datos", Ra-Ma, Madrid.

Miles, A. 2006, *A Thesaurus Data Model for British Standard 8723*.

Miles, A. 2003 "RDF Molecules: Evaluating Semantic Web Technology in a Scientific Application" *Poster submission to WWW Conference 2003 on performance of RDF toolkits (Jena)* Disponible en: http://www.w3c.rl.ac.uk/SWAD/papers/RDFMolecules_final.doc

Miles, A. y Bechhofer, S. 2008, W3C Working Draft 29 August 2008, *SKOS Simple Knowledge Organization System* Disponible en: http://www.w3.org/TR/skos-reference/.

Miles, A., Matthews, B., Wilson, M. y Brickley, D. 2005, "SKOS core: simple knowledge organisation for the web", *Proceedings of the 2005 international Conference on Dublin Core and Metadata Applications: Vocabularies in Practice*Universidad Carlos III, Madrid.

Miller, G.A. 1995, "WordNet: a lexical database for English", *Communications of the ACM,* vol. 38, no. 11, pp. 39-41 Disponible en: http://doi.acm.org/10.1145/219717.219748; http://delivery.acm.org/10.1145/220000/219748/p39-miller.pdf?key1=219748ykey2=3007292521ycoll=GUIDEydl=GUIDEyCFID=51659862yCFTOKEN=85134438.

Milstead, J.L. 2000*, About Thesauri*. En Bayside Indexing Service. Disponible en: http://www.bayside-indexing.com/Milstead/about.htm.

Mochón Bezares, G. y Sorli Rojo, Á. 2007, "Tesauros de ciencias sociales en internet", *Revista Española de Documentación Científica,* vol. 30, no. 3, pp. 395-419 Disponible en: http://redc.revistas.csic.es/index.php/redc/article/viewFile/392/404.

Moliner, M. 1998, "Diccionario de uso del español", Gredos, Madrid.

Montejo Ráez, A. 2001, "Proyecto de indexado automático para documentos en el campo de la física de altas energías", *Procesamiento del lenguaje natural,*no. 27, pp. 295-296 Disponible en: http://rua.ua.es/dspace/handle/10045/1824; http://rua.ua.es/dspace/bitstream/10045/1824/1/PLN_27_36.pdf.

Monti Bonafede, S., San Vicente, F. y Preti, V. 2006, "Characteristics and Capacity of e-learning platforms for learning languages", *eLearning Papers,* vol. 1 Disponible en: http://www.elearningpapers.eu/index.php; http://www.elearningpapers.eu/index.php?page=docydoc_id=8401ydoclng=6.

Moreiro González, J.A. 2006, "La representación y recuperación de los contenidos digitales: de los tesauros conceptuales a las folksonomías" in *Tendencias en documentación digital*, ed. J. Tramullas Saz, Trea, , pp. 81-109.

Moya, G., Gil, I. 2001, "Evaluación de softwares de gestión de tesauros", *Ciencias de la Información* Vol. 32, No. 3, diciembre, 2001, Disponible en: http://webs.um.es/isgil/Software%20gestion%20tesauros%20MOYA%20MARTINEZ%20y%20GIL-LEIVA,%20Isidoro.pdf

Murray, T. 1999, "Authoring Intelligent Tutoring Systems: An analysis of the state of the art", *International Journal of Artificial Intelligence in Education,* vol. 10, pp. 98-129 Disponible en: http://ihelp.usask.ca/iaied/ijaied/members99/archive/vol_10/murray/full.html.

Nakayama, K., Hara, T. y Nishio, S. 2007, "A Thesaurus Construction Method from Large ScaleWeb Dictionaries", *Proceedings of the 21st international Conference on Advanced Networking and Applications*AINA. IEEE Computer Society, Washington, DC, pp. 932 Disponible en: http://dx.doi.org/10.1109/AINA.2007.23.

NAL Thesaurus, 2008, Tesauro de la Biblioteca Nacional de Agricultura de EEUU, 2008th edn Disponible en: http://agclass.nal.usda.gov/agt.shtml, http://agclass.nal.usda.gov/agt_es.shtml.

Neven, F. y Duval, E. 2002, "Reusable learning objects: a survey of LOM-based repositories", ACM, New York, USA, pp. 291-294 Disponible en: http://delivery.acm.org/10.1145/650000/641067/p291-neven.pdf?key1=641067ykey2=4526602521ycoll=GUIDEydl=GUIDEyCFID=50190701yCFTOKEN=11283208.

Nilsson, M., Baker, T. y Johnston, P., 2009, "Interoperability Levels for Dublin Core Metadata", DCMI Recommended Resource, Identifier: http://dublincore.org/documents/2009/05/01/interoperability-levels/ Disponible en: http://dublincore.org/documents/interoperability-levels/

Novak, J.D. y Cañas, A.J. 2008, "The Theory Underlying Concept Maps and How to Construct and Use Them", Technical Report IHMC CmapTools 2006-01 Rev 2008-01, Institute for Human and Machine Cognition, Florida Disponible en: http://cmap.ihmc.us/Publications/ResearchPapers/TheoryCmaps/TheoryUnderlyingConceptMaps.htm

Noy, N. y Rector, A. (eds) 2006, "Defining N-ary Relations on the Semantic Web", *W3C Working Group Note 12 April 2006*. Disponible en: http://www.w3.org/TR/swbp-n-aryRelations/

Nyhan, J. 2008, "Developing Integrated Editions of Minority Language Dictionaries: The Irish Example", *Literary and Linguistic Computing,* vol. 23, no. 1, pp. 3-12

Disponible en: http://llc.oxfordjournals.org/cgi/reprint/23/1/3; http://llc.oxfordjournals.org/cgi/content/full/23/1/3.

Or-Bach, R. 2005, "Educational benefits of metadata creation by students", *ACM SIGCSE Bulletin,* vol. 37, no. 4, pp. 93 Disponible en: http://doi.acm.org/10.1145/1113847.1113885; http://delivery.acm.org/10.1145/1120000/1113885/p93-or-bach.pdf?key1=1113885ykey2=5049292521ycoll=GUIDEydl=GUIDEyCFID=51664820yCFTOKEN=36473766.

Paige, R.F. 1995, *Higraph-Based Predicate and Heterogeneous Specification* Disponible en: http://www.cs.utoronto.ca/; http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.848.

Panizo, L.; Sánchez, L.; Fernández, B. y Llamas, M. (eds.), 2006, *Proceedings of the 8th International Symposium on Computers in Education*, Universidad de León, León.

Pastor, J.A. y Martínez, F.J. 2003, "iSGAT: Gestión colaborativa de tesauros en Internet", *Scire: Representación y organización del conocimiento,* vol. 9, no. 2, pp. 85-98 Disponible en: http://www.um.es/gtiweb/fjmm/isgat.pdf; http://ibersid.eu/ojs/index.php/scire/article/viewFile/1468/1446.

Paynter, G.W., Witten, I.H., Cunningham, S.J. y and Buchanan, G. 2000, "Scalable browsing for large collections: a case study", *Proceedings of the Fifth ACM Conference on Digital Libraries*ACM, New York, NY, US, pp. 215 Disponible en: http://doi.acm.org/10.1145/336597.336666.

Pérez Agüera, J.R. 2004, *Automatización de tesauros y su utilización en la web semántica*. *BiD: textos universitaris de biblioteconomia i documentació*, núm. 13 (dic. 2004). Universidad de Barcelona. Facultad de Biblioteconomía y DocumentaciónDisponible en: http://www.w3.org/2001/sw/Europe/events/200406-esp/trabajo-final-extratesauros/trabajo-final-extratesauros.html.

Pidcock, W. 2003, *What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology and a meta-model?*. Disponible en: http://www.metamodel.com/article.php?story=20030115211223271.

PLS Ramboll, 2004, *Studies in the Context of the E-learning Initiative: Virtual Models of European Universities,* Management for the European Commission Disponible en: http://www.elearningeuropa.info/extras/pdf/virtual_models.pdf.

Pollard, R. 1993, "A hypertext-based thesaurus as a subject browsing aid for bibliographic databases", *Information Processing y Management,* vol. 29, no. 3, pp. 345-357 Disponible en: http://dx.doi.org/10.1016/0306-4573(93)90060-Q.

Powell, A., Nillson, M., Naeve, A., Jonhston, P. y Baker, T. 2005,*DCMI Abstract Model*. Disponible en: http://dublincore.org/documents/abstract-model/.

Protopsaltis, A. y Bouki , V. 2005, "Towards a hypertext reading/comprehension model", *Proceedings of the 23rd annual international conference on Design of communication: documenting y designing for pervasive information*, eds. S. Tilley y R. Newman, ACM Press, New York, NY, USA, pp. 159 Disponible en: http://0-doi.acm.org.cisne.sim.ucm.es:80/10.1145/1085313.1085349.

Quillan, R. 1968, "Semantic memory" in *Semantic Information Processing*, ed. M. Minsky, M.I.T. Press, Cambridge, Mass., USA, pp. 227-268.

Rada, R. y Martin, B.K. 1987, "Augmenting thesauri for information systems", *ACM Transactions on Information Systems,* vol. 5, no. 4, pp. 378-392 Disponible en: http://doi.acm.org/10.1145/42196.42246.

RAE 2001, "Diccionario de la lengua española / Real Academia Española", Espasa, Madrid Disponible en: http://buscon.rae.es/draeI.

Ramón y Cajal, S. 1941, "Reglas y consejos sobre investigación científica. Los tónicos de la voluntad", Espasa Calpe, Madrid.

Ranaganathan, S.R. 1987, "The Colon classification", Sarada Ranganathan Endowment for Library Science, Bangalore.

RDF, 2004, REC-rdf-concepts-20040210, *Resource Description Framework (RDF): Concepts and Abstract Syntax*, W3C Recommendation 10 February 2004, Klyne, G. y Carroll, J.J. (Eds.) Disponible en: http://www.w3.org/TR/rdf-concepts/

RDFS, 2004, *RDF Vocabulary Description Language 1.0: RDF Schema*, W3C Recommendation 10 February 2004, Brickley, D; Guha, R.V. (Eds.) Disponible en: http://www.w3.org/TR/rdf-schema/.

Rodríguez, J.V. 1997, "Un modelo de datos para la construcción de tesauros", *Investigación bibliotecológica,* vol. 11, no. 22, pp. 39-50 Disponible en: http://www.ejournal.unam.mx/ibi/vol11-22/IBI001102204.pdf.

Rodríguez Perojo, K. y Ronda León, R. 2005, "Web semántica: un nuevo enfoque para la organización y recuperación de información", *ACIMED,* vol. 13, no. 6 Disponible en: http://bvs.sld.cu/revistas/aci/vol13_6_05/aci03605.htm.

Roget, P.M. 1852, "Roget's Thesaurus of English Words and Phrases", 2002, ed. G. Davidson, Penguin.

Romiszowski, A. 2004, "How's the E-learning Baby? Factors Leading to Success or Failure of an Educational Technology Innovation", *Educational Technology,* vol. 44,

no. 1, pp. 5-27 Disponible en: http://www.elearning-reviews.org/topics/resources-management/project-management/2004-romiszowski-elearning-baby/; http://www.itslifejimbutnotasweknowit.org.uk/files/elearning_failure_study-romiszowsky.pdf.

Salton, G. y McGill, M.J. 1986, "Introduction to Modern Information Retrieval", McGraw-Hill Book Company, New York.

Sampson, D.G., Lytras, M.D., Wagner, G. y Díaz, P. 2004, "Ontologies and the Semantic Web for E-learning", *Educational Technology & Society,* vol. 7, no. 4, pp. 26-28. Disponible en: http://www.ifets.info/journals/7_4/5.pdf

Sanz, A., y Fernández-Pampillón, A. (2009). "Un plan estratégico para la calidad en la Facultad de Filología". V Jornadas de Campus Virtual UCM, Madrid, 10 de Febrero de 2009 (pendiente de publicación e-prints UCM).

Santanach Delisau, F., Casamajó Dalmau, J., Casado Arias, P. y Alier Forment, M. 2007, "Proyecto CAMPUS. Una plataforma de integración", *IV Simposio Pluridisciplinar sobre Diseño, Evaluación y Desarrollo de Contenidos Educativos Reutilizables*, eds. M. Benito, J. Romo y J. Portillo, Disponible en: http://spdece07.ehu.es/actas/Santanach.pdf.

Sarasa Cabezuelo, A., Canabal, J.M., Sacristán, J.C. y Jiménez, R. 2008, "Uso de IMS VDEX en Agrega", *X Simposio Internacional de Informática Educativa SIIE 2008*Universidad de Salamanca, Salmnca, España, pp. 119.

Saussure, F.de. 1916, "Curso de Lingüística General", Editorial Losada, Buenos Aires (2007)

Schaffert, S. y Hilzensauer, W. 2008, "On the way towards Personal Learning Environments: Seven crucial aspects", *eLearning Papers,* vol. 9 Disponible en: http://www.elearningpapers.eu; http://www.elearningeuropa.info/files/media/media15971.pdf.

Schank, R.C. y Tesler, L. 1969, "A conceptual dependency parser for natural language", *Proceedings of the 1969 conference on Computational linguistics*ACM, Morristown, NJ, USA.

Schauble, P. 1987, "Thesaurus based concept spaces", *Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information retrieval*, eds. C.T. Yu y C.J.v. Rijsbergen, ACM Press, New York, NY, USA, pp. 254.

SCORM 2004, *Sharable Content Object Reference Model (SCORM) 2004 3rd Edition Documentation Suite* Disponible en: http://www.adlnet.gov/Downloads/DownloadPage.aspx?ID=237;

http://www.adlnet.gov/Technologies/scorm/SCORMSDocuments/Forms/AllItems.aspx?RootFolder=%2fTechnologies%2fscorm%2fSCORMSDocuments%2fPrevious%20Versions%2fSCORM%202004%203rd%20EdyFolderCTID=0x0120007F801FCD5325044C89D91240519482D7yView={4D6DFFDE-3CFC-4DD9-A21A-4B687728824A}.

Seco, M., Andrés, O. y Ramos, G. 1999, "Diccionario del Español Actual", Aguilar, Madrid.

Sérasset, G., Brunet-Manquat, F. y Chiocchetti, E. 2006, "Multilingual Legal Terminology on the Jibiki Platform: The LexALP Project", *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*Association for Computational Linguistics, pp. 937 Disponible en: http://www.aclweb.org/anthology/P/P06/P06-1118.pdf.

Shapiro, C.D. y Yan, P. 1996, "Generous Tools: Thesauri in Digital Libraries", *17th National Online Meeting Proceedings* Information Today, Medford, NJ.

Shieber, S.M. 1986, "An Introduction to Unification-Based Approaches to Grammar", Center for the Study of Language and Information, Leland Stanford Junior University, US. Versión en español 1989 "Introducción a los formalismos gramaticales de unificación". Editorial Teide

Sierra, J.L. y Fernández-Valmayor, A. 2006, " A Heritage Dissemination Approach for the Production and Maintenance of Repositories of Learning Objects", *Proceedings of the VIII Simposio Internacional de Informática Educativa* Disponible en: http://www.e-ucm.es/drafts/e-UCM_draft_57.pdf.

Sierra, J.L., Fernández-Valmayor, A., Guinea Bueno, M., Hernanz, H. y Navarro, A. 2005, "Building Repositories of Learning Objects in Specialized Domains: The Chasqui Approach", *Proceedings of the 5th IEEE International Conference on Advanced Learning Technologies*IEEE Computer Society, pp. 225 Disponible en: http://www2.computer.org/portal/web/csdl/doi/10.1109/ICALT.2005.77.

Sini, M., Lauser, B., Salokhe, G., Keizer, J. y Katz, S. 2008, "The AGROVOC Concept Server: rationale, goals and usage", *Library Review,* vol. 57, no. 3, pp. 200-212 Disponible en: ftp://ftp.fao.org/docrep/fao/010/ai167e/ai167e00.pdf.

Sioutos, N., Coronado, S.d., Haber, M.W., Hartel, F.W., Shaiu, W. y Wright, L.W. 2007, "NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information", *Journal of Biomedical Informatics,* vol. 40, no. 1, pp. 30-43 Disponible en: http://dx.doi.org/10.1016/j.jbi.2006.02.013.

Slavic, A. 2000, *A Definition of Thesauri and Classification as Indexing Tools*. Disponible en: http://dublincore.org/documents/2000/11/28/thesauri-definition/.

Slype, G.v. 1991, "Los lenguajes de indización: Concepción, construcción y utilización en los sistemas documentales", Fundación Germán Sánchez Ruipérez, Pirámide, D.L., Madrid.

Soergel, D. 2002, "Thesauri and ontologies in digital libraries: 1. structure and use in knowledge-based assistance to users", *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*ACM, New York, NY, USA, pp. 419 Disponible en: http://delivery.acm.org/10.1145/550000/544364/p415-soergel.pdf?key1=544364ykey2=3327332521ycoll=GUIDEydl=GUIDEyCFID=51771822yCFTOKEN=53825875.

Soergel, D. 2002, "Thesauri and ontologies in digital libraries: 2. design, evaluation, and development", *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*ACM, New York, NY, USA, pp. 419 Disponible en: http://portal.acm.org/citation.cfm?doid=544220.544368; http://www.dsoergel.com/cv/B63.pdf; http://delivery.acm.org/10.1145/550000/544368/p419-soergel.pdf?key1=544368ykey2=9808952521ycoll=GUIDEydl=GUIDEyCFID=52252481yCFTOKEN=94918741.

Sowa, J.F. 2000, "Knowledge Representation: Logical, Philosophical, and Computational Foundations", Brooks Cole Publishing Co., Pacific Grove, CA, pp. 594 Disponible en: http://www.jfsowa.com/krbook/.

Sowa, J.F. 2000b, "Ontology, Metadata, and Semiotics", in B. Ganter & G. W. Mineau, eds., *Conceptual Structures: Logical, Linguistic, and Computational Issues*, Lecture Notes in AI #1867, Springer-Verlag, Berlin, 2000, pp. 55-81. Disponible en: http://users.bestweb.net/~sowa/peirce/ontometa.htm

Sowa, J.F. 2005, "Concepts in the Lexicon: Introduction", Last Modified: 11/27/2005, Disponible en: http://www.jfsowa.com/ontology/lexicon.htm

Sowa, J.F. (ed.), 1991, "Principles of Semantic Networks: Explorations in the Representation of Knowledge", Morgan Kaufmann Publishers, San Marco, CA.

SPINES thesaurus Disponible en: http://pci204.cindoc.csic.es/tesauros/SpinTes/HTML/SPI_A4.htm.

SQL Language Reference 2008, 11g Release 1 B28286 Mayo 2008 Disponible en:http://download.oracle.com/docs/cd/B28359_01/server.111/b28286.pdf

Staab, S. y Studer, R. (eds.), 2004, "Handbook on Ontologies", Springer.

Stewart, I. 2004, "De aquí al infinito. Las matemáticas de hoy", Editorial Crítica.

Stuckenschmidt, H. y Harmelen, F.v. 2005, "Information Sharing on the Semantic Web", Springer, Berlin, pp. 276.

Tabata, K. y Mitsumori, S. 2002, "An Assertion-Based Information-Probe System: Document-Skeleton and Glossary-Skeleton Approach", *Information-Knowledge-Systems Management,* vol. 3, no. 2-4, pp. 123-152 Disponible en: http://portal.acm.org/citation.cfm?id=946022.

Taghva, K., Borsack, J. y Condit, A. 1999, "The Efectiveness of Thesauri-Aided Retrieval", *Proceedings of the ISyT/ SPIE Symposium on Electronic Imaging: Science and Technology* Disponible en: http://www.isri.unlv.edu/publications/isripub/Taghva99a.pdf.

Taylor, A.G. 2006, "Introduction to Cataloging and Classification", Libraries Unlimited, Englewood, Colorado.

TEI-P5, 2007, *P5: Guidelines for Electronic Text Encoding and Interchange - v. A gentle introduction to XML* Disponible en: http://www.tei-c.org/release/doc/tei-p5-doc/html/SG.html.

Teorey, T.J., Yang, D. y Fry, J.P. 1986, "A logical design methodology for relational databases using the extended entity-relationship model", *ACM Computing Surveys,* vol. 18, no. 2, pp. 197-222 Disponible en: http://doi.acm.org/10.1145/7474.7475.

Tesauro AGROVOC Disponible en: http://www.fao.org/aims/ag_intro.htm.

Tesauro EUROVOC. Tesauro de la Unión Europea Disponible en: http://www.r020.com.ar/eurovoc/index.php.

Tesauro de Educación Superior 1998, Disponible en: http://www.termilat.info/public/env746.dot.

Thayer, R.H. 2003, "Software Engineering Glossary", *IEEE Software,* vol. 20, no. 4.

Tompa, F.W. 1989, "A data model for flexible hypertext database systems", *ACM Transactions on Information Systems,* vol. 7, no. 1, pp. 85-100 Disponible en: http://doi.acm.org/10.1145/64789.64993.

Trigari, M., 2002, ETB THESAURUS - Description and Comments - European Schoolnet, Web Editor: Riina Vuorikari, Published: Thursday, 21 Feb 2002, Last changed: Wednesday, 18 Aug 2004, Disponible en: http://etb.eun.org/eun.org2/eun/en/etb/content.cfm?lang=en&ov=12233

Tudhope, D., Alani, H. y Jones, C. 2001, "Augmenting Thesaurus Relationships: Possibilities for Retrieval", *Journal of Digital Information,* vol. 1, no. 8 Disponible en: http://jodi.tamu.edu/Articles/v01/i08/Tudhope/.

Ullman, J.D. 1988, "Principles of Database and Knowledge-Base System", Computer Science Press, Rockville, Md.

UMLS, 2008, Unified Medical Language System, **Last updated:** 02 April 2009, Disponible en: http://semanticnetwork.nlm.nih.gov/.

UMLS, 2009, April Release AA2009, *UMLS Knowledge Source - Documentation*, U.S. National Library of Medicine, Bethesda (MD) Disponible en: http://www.nlm.nih.gov/research/umls/umlsdoc.html.

UNE 50106, 1990 y 1995, *Directrices para el establecimiento y desarrollo de tesauros monolingües*.

Unesco Thesaurus Disponible en: http://www.ulcc.ac.uk/unesco/.

Van Heijst, G.v., Schreiber, A.T. y Wielinga, B.J. 1997, "Using explicit ontologies in KBS development", *International Journal of Human-Computer Studies,* vol. 46, no. 2-3, pp. 183-292 Disponible en: http://portal.acm.org/citation.cfm?id=250542; http://www.sciencedirect.com/science?_ob=ArticleURLy_udi=B6WGR-45M91MP-Vy_user=144492y_rdoc=1y_fmt=y_orig=searchy_sort=dy_docanchor=yview=cy_acct=C000012038y_version=1y_urlVersion=0y_userid=144492ymd5=3abac94b8530450af53acb2eceafa6df.

Van-Dusen, G.C. 1997, "The Virtual Campus. Technology and Reform in Higher Education", *ASHE-ERIC Higher Education Report,* vol. 25, no. 5.

VDEX–IMS 2004, *IMS Vocabulary Definition Exchange (VDEX) specification* Disponible en: http://www.imsglobal.org/vdex/index.html.

Velardi, P., Navigli, R. y D'Amadio, P. 2008, "Mining the Web to Create Specialized Glossaries", *IEEE Intellingent Systems,* vol. 23, no. 5, pp. 18-25 Disponible en: http://doi.ieeecomputersociety.org/10.1109/MIS.2008.88; http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=yarnumber=4629722yisnumber=4629713?tag=1.

VisualThesaurus Disponible en: www.visualthesaurus.com.

Walker, D.E., Zampolli, A. y Calzolari, N. 1995, "Automating the Lexicon: Research and Practice in a Multilingual Environment" in Oxford University Press.

Wang, W. y Randa, R. 1998, "Structured hypertext with domain semantics", *ACM Transactions on Information Systems,* vol. 16, no. 4, pp. 372-412 Disponible en: http://0-doi.acm.org.cisne.sim.ucm.es:80/10.1145/291128.291132.

Wilson, M. y Matthews, B. 2002, "Migrating Thesauri to the Semantic Web", *European Research Consortium for Informatics and Mathematics,* vol. 51 Disponible en: http://www.ercim.org/publication/Ercim_News/enw51/EN51.pdf.

Woods, W.A., Bookman, L.A., Houston, A., Kuhns, R.J., Martin, P. y Green, S. 2000, "Linguistic Knowledge Can Improve Information Retrieval", *Proceedings of ANLP-2000, April 29 - May 4, Seattle, Washington, USA* Disponible en: http://research.sun.com/features/tenyears/volcd/papers/20Woods.pdf.

XML, 2008, REC-xml-20081126, "Extensible Markup Language (XML) 1.0 (Fifth Edition)", W3C Recommendation 26 November 2008, Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F. (Eds.) Disponible en: http://www.w3.org/TR/REC-xml/

Yan, P., Jiao, Y., Hurson, A.R. y Potok, T.E. 2006, "Semantic-based information retrieval of biomedical data", *Proceedings of the 2006 ACM symposium on Applied computing*ACM Press, New York, NY, USA, pp. 1700 Disponible en: http://doi.acm.org/10.1145/1141277.1141678.

Yang, D. y Powers, D. M. 2008. "Automatic thesaurus construction". In *Proceedings of the Thirty-First Australasian Conference on Computer Science - Volume 74* (Wollongong, Australia, January 01 - 01, 2008). G. Dobbie and B. Mans, Eds. ACSC, vol. 312. Australian Computer Society, Darlinghurst, Australia, 147-156.

Zadrony, W. 1991, "Logical dimensions of some graph formalisms" in *Principles of semantic networks. Explorations in the representation knowledge*, ed. J.F. Sowa, Morgan Kaufmann Publishers, Inc., San Mateo, California.

Zernik, U. (ed.), 1991, "Lexical Acquisition. Exploiting on-line resources to build a lexicon", Lawrence Erlbaum Associates, Inc. Publishers.

Apéndice A. Índice de tesauros

NOMBRE	SIGLA	AUTOR	LOCALIZACIÓN
Tesauro Agrícola/ Agricultural thesaurus		Biblioteca Nacional de Agricultura de EE.UU. (NAL por sus siglas en inglés) y el Instituto Interamericano de Cooperación para la Agricultura (IICA)	http://agclass.canr.msu.edu/agt_es.shtml/ http://agclass.nal.usda.gov/
AGROVOC Thesaurus	AGROVOC	AGRIS. CARIS. FAO	http://aims.fao.org/website/AGROVOC-Thesaurus/sub
ARIADNE thesaurus	ARIADNE thesaurus	ARIADNE foundation for European Knowledge pool	http://ariadne.cs.kuleuven.be/silo2006/visualbrowse.jsp; www.**ariadne**-eu.org/
Art and Architecture Thesaurus	AAT	The Getty	http://www.getty.edu/research/conducting_research/vocabularies/aat/#sample
CERES thesaurus	CERES	The California Environmental Resources Evaluation System	http://ceres.ca.gov/thesaurus/Overview.htm, http://ceres.ca.gov/education/
Cyc ontology		Cycorp	http://taxonomies.cyc.com/
Dewey Decimal Classification	DDC	M. Dewey	http://www.oclc.org/oclc/fp/
European Language Social Science Thesaurus	ELSST	Council of European Social Science Data Archives	http://www.w3c.rl.ac.uk/pasttalks/slidemaker/XML_UK_SW_Thes/ELSST_economics.xml; http://www.w3c.rl.ac.uk/SWAD/thes_links.htm
ERIC Thesaurus	ERIC	Educational Resource Information Center	www.eric.ed.gov/thesaurus
European Tresaury Browser	ETB	European Commission, Bruselas	http://etb.eun.org/etb/index.htm
EUROVOC Thesaurus	EUROVOC	European Communities	http://europa.eu/eurovoc/; http://www.r020.com.ar/eurovoc/index.php
Gateway to Educational Materials	GEM	Gateway	www.thegateway.org
IEDCYT Tesauros	IEDCYT (antes CINDOC)	Instituto de Estudios Documentales sobre Ciencia y Tecnología	http://thes.cindoc.csic.es/index_esp.html
Library of Congress Classification	LCC	Library of Congress (Estados Unidos)	http://lcweb.loc.gov

Library of Congress Subject Headings	LCSH	Library of Congress (Estados Unidos)	http://lcweb.loc.gov
Language Resource Exchange thesaurus	LRE	European Schoolnet	http://fire.eun.org/70.xml
Macrotesauro Mexicano para Contenidos Educativos	MMCE	CUIB	http://cuib.unam.mx/~tesauro
Medical Subject Headings	MeSH	National Library of Medicine	http://www.nlm.nih.gov/mesh/2007/index.html; http://www.nlm.nih.gov/mesh/2009/mesh_browser/MeSHtree.I.html
Multimedia Educational Resource for Learning and Online Teaching thesaurus	MERLOT thesaurus	MERLOT	http://www.merlot.org/merlot/index.htm
Multiwordnet		Varios grupos de investigación	http://multiwordnet.itc.it/online/multiwordnet.php
National Agricultural Library Thesaurus	NAL	National Agricultural Library	http://agclass.nal.usda.gov/agt.shtml (inglés), http://agclass.nal.usda.gov/agt_es.shtml (español)
Sistema de Intercambio de Información sobre Políticas Científicas y Tecnológicas	SPINES	UNESCO e ICYT (IEDCYT)	http://thes.cindoc.csic.es/index_SPIN_esp.html; http://spines.r020.com.ar/index.php; http://pci204.cindoc.csic.es/tesauros/SpinTes/HTML/SPI_A4.htm
Tesauro de la Biblioteca de la UCM		UCM	http://alfama.sim.ucm.es/greco/t-digital.php
Tesauro de Biblioteconomía y Documentación		IEDCYT	http://thes.cindoc.csic.es/index_BIBLIO_esp.html
Tesauros de ciencias sociales en internet		Mochón Bezares, G. y Sorli Rojo, Á.	http://redc.revistas.csic.es/index.php/redc/article/viewFile/392/404
Tesauro de Derecho		IEDCYT	http://thes.cindoc.csic.es/index_DEREC_esp.html
Tesauro Europeo de la Educación	TEE		http://www.freethesaurus.info/redined/es/index.php; http://redined.r020.com.ar/es/
Tesauro de Educación Superior	TES	UCM	http://www.terminometro.info/ancien/b37/es/tesauro_es.htm; http://www.termilat.info/public/env746.dot
Thesaurus for Graphic Materials	TGM	Prints and Photographs Division Library of Congress	http://www.loc.gov/rr/print/tgm1/toc.html
tesauro INSPEC	INSPEC	The Institution of Electrical Engineers	http://www.cs.ucla.edu/Leap/Eisa/inspec.htm#thesaurus
Tesauro de la NASA		NASA	http://www.sti.nasa.gov/products.html#pubtools
Tesauro OIT		Organización Internacional del trabajo	http://www.ilo.org/public/libdoc/ILO-Thesaurus/spanish/

Tesauro de Propiedad Industrial		IEDCYT	http://thes.cindoc.csic.es/index_PROIND_esp.html
Tesauro de Psicología		IEDCYT	http://thes.cindoc.csic.es/index_PSICO_esp.html
Tesauro de Redes de Ordenadores		Martínez, F.J. y García, J.C.	http://www.um.es/gtiweb/fjmm/tesauro/index.html
Tesauro de Urbanismo		IEDCYT	http://thes.cindoc.csic.es/index_URBA_esp.html
Lista de Tesauros en XML			http://www.w3c.rl.ac.uk/SWAD/thes_links.htm
Thesaurus of Engineering and Scientific Terms	TEST	Committee on Scientific and Technical Information	
Thesaurus Maths			http://www.thesaurus.maths.org
Universal Decimal Classification	UDC	P. Otlet y H. La Fontaine	http://zeus.slais.ucl.ac.uk/udc/
UNESCO Thesaurus (Tesauro de la UNESCO)		UNESCO	http://databases.unesco.org/thesaurus/, http://www2.ulcc.ac.uk/unesco/
Unified Medical Language System	UMLS	US National Library of Medicine	http://www.nlm.nih.gov/research/umls/; http://semanticnetwork.nlm.nih.gov/
VisualThesaurus		Thinkmap, Inc.	www.visualthesaurus.com
Wordnet		Cognitive Science Laboratory Universidad de Princeton George A. Miller	http://www. wordnet.princeton.edu; http://www.visuwords.com/

Apéndice B. Esquema relacional SQL de un Higraph Léxico[1]

```
CREATE TABLE  "HIGRAPH_MICRO_RELS"
   (    "ID" VARCHAR2(255) NOT NULL ENABLE,
        "TYPE" VARCHAR2(255),
        "INCONSISTENCY" VARCHAR2(1),
         CONSTRAINT "MICRO_RELS_NODETYPE_CK" CHECK (type in ('category'
,'term' ,'relation')) ENABLE,
         CONSTRAINT "MICRO_RELS_INCONSISTENCY_CK" CHECK (inconsistency
in ('N' ,'S' )) ENABLE,
         CONSTRAINT "MICRO_RELS_PK" PRIMARY KEY ("ID") ENABLE
   )


CREATE TABLE  "HIGRAPH_MACRO_RELS"
   (    "REL_ID" VARCHAR2(255) NOT NULL ENABLE,
        "NODE_ID1" VARCHAR2(255),
        "NODE_ID2" VARCHAR2(255),
        "FRECUENCY" NUMBER(5,0),
        "INCONSISTENCY" VARCHAR2(1),
         CONSTRAINT "MACRO_RELS_INCONSISTENCY_CK" CHECK (inconsistency
in ('N' ,'S' )) ENABLE,
         CONSTRAINT "MACRO_RELS_TYPE_FK" FOREIGN KEY ("REL_ID")
          REFERENCES  "HIGRAPH_REL_TYPES" ("ID") ENABLE,
         CONSTRAINT "MACRO_RELS_NODO1_FK" FOREIGN KEY ("NODE_ID1")
          REFERENCES  "HIGRAPH_MICRO_RELS" ("ID") ENABLE,
         CONSTRAINT "MACRO_RELS_NODO2_FK" FOREIGN KEY ("NODE_ID2")
          REFERENCES  "HIGRAPH_MICRO_RELS" ("ID") ENABLE
   )


CREATE TABLE  "HIGRAPH_REL_TYPES"
   (    "ID" VARCHAR2(255) NOT NULL ENABLE,
        "TYPE" VARCHAR2(255),
         CONSTRAINT "REL_TYPES_CK" CHECK (type in
('orden','equivalencia')) ENABLE,
         CONSTRAINT "REL_TYPES_PK" PRIMARY KEY ("ID") ENABLE
   )
```

[1] Creado con ORACLE Database Express Edition.